I

INTERPOLATION DE LAGRANGE

§ 1 INTRODUCTION À L'INTERPOLATION POLYNOMIALE

1.1 Espaces de polynômes

Nous rappelons quelques résultats sur les polynômes (ou fonctions polynomiales). Un **monôme** de degré k est une fonction de la forme $x \in \mathbb{R} \to cx^k$ où $c \in \mathbb{R}^\star$ et $k \in \mathbb{N}$. Un **polynôme** est une somme (finie) de monômes. La fonction nulle est aussi considérée comme un polynôme. L'ensemble $\mathscr{P}$ des polynômes forme alors un espace vectoriel quand on utilise l'addition habituelle des fonctions $(p+q)$ ainsi que la multiplication par une constante (λp). Le produit de deux polynômes (pq) est encore un polynôme. Les fonctions polynômes sont indéfiniment dérivables. Tout polynôme *p non nul* s'écrit d'une manière et d'une seule sous la forme

$$p(x) = c_0 + c_1 x + \cdots + c_m x^m = \sum_{i=0}^{m} c_i x^i \tag{1.1}$$

avec $c_m \neq 0$. L'unicité provient de ce que $c_k = p^{(k)}(0)/k!$. Les nombres c_i s'appellent les coefficients de p. L'entier non nul m dans (1.1) est le **degré** de p et le coefficient c_m est le **coefficient dominant** de p. On convient que $\deg 0 = -\infty$. Avec cette convention, quels que soient les polynômes p et q on a

$$\deg(pq) = \deg p + \deg q \tag{1.2}$$

$$\deg(p+q) \leq \max(\deg p, \deg q). \tag{1.3}$$

E. 1. Ecrire une formule donnant les coefficients d'un produit de polynômes pq en fonction des coefficients des facteurs p et q.

Lorsque $\lambda \in \mathbb{R}^\star$,

$$\deg \lambda p = \deg p, \tag{1.4}$$

c'est un cas particulier de (1.2). En réalité le degré de $p+q$ coïncide toujours avec $\max(\deg p, \deg q)$ sauf lorsque les deux polynômes ont même degré et leurs coefficients dominants sont opposés l'un de l'autre. On note $\mathscr{P}_m$ l'ensemble des polynômes de degré inférieur ou égal à m. Les propriétés (1.3) et (1.4) montrent que $\mathscr{P}_m$ est un sous-espace vectoriel dont la base canonique est $\mathscr{B} = (x \to x^0, x \to x^1, \ldots, x \to x^m)$. En particulier sa dimension est $m+1$.

Si r est une racine de p (c'est-à-dire $p(r) = 0$) alors p est divisible par $(\cdot - r)$. Cela signifie qu'il existe un polynôme q tel que $p(x) = (x-r)q(x)$ pour tout $x \in \mathbb{R}$. On dit que r est une racine de **multiplicité** m lorsque $(\cdot - r)^m$ divise p mais $(\cdot - r)^{m+1}$ ne divise pas p. On montre en algèbre que cela est équivalent à

$$0 = p(r) = p'(r) = \cdots = p^{(m-1)}(r) \quad \text{et} \quad p^{(m)}(r) \neq 0.$$

Un polynôme $p \in \mathscr{P}_m$ *non nul* admet au plus m racines en tenant compte de la multiplicité. Cela signifie que si r_i est racine de multiplicité m_i de $p \neq 0$ pour $i = 1, \ldots, l$ alors $m_1 + \cdots + m_l \leq m$. On dit alors que le nombre de racine de p est en tenant compte de la multiplicité plus petite ou égale au degré du polynôme p*. On utilisera plusieurs fois que si p est un polynôme de degré au plus m qui admet au moins $m+1$ racines en tenant compte de la multiplicité alors p est nécessairement le polynôme nul, autrement dit

$$\left.\begin{array}{l} z_i \text{ racine de } p \text{ de multiplicité} \geq m_i,\ i = 1, \ldots, l, \\ \sum_{i=1}^{l} m_i > m, \\ p \in \mathscr{P}_m \end{array}\right\} \implies p = 0.$$

E. 2. Peut-on retrouver un polynôme quand on connaît toutes ses racines?

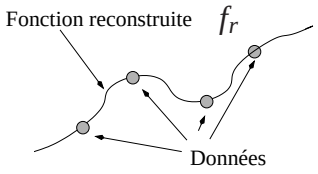
1.2 Construction de l'interpolant de Lagrange

a) Le problème général de l'interpolation polynomiale

En analyse numérique, une fonction f n'est souvent connue que par ses valeurs f_i en un nombre fini de points a_i, $f_i = f(a_i)$, (en réalité, en pratique f_i est seulement une approximation de $f(a_i)$). Cependant, dans la plupart des cas, il

* Dans le cas complexe, c'est-à-dire, lorsque'on accepte de considérer les racines complexes (et mêmes les polynômes à coefficients complexes), le théorème fondamental de l'algèbre dit que le nombre de racines d'un polynôme non nul est, en tenant compte de la multiplicité, exactement égal au degré du polynôme.

est nécessaire d'effectuer des opérations sur des fonctions *globales* (dérivation, intégration, ...) et on est donc conduit à *reconstruire* une fonction globale f_r à partir d'un nombre fini de données (a_i, f_i).



Sauf cas très simple, la fonction f_r ne coïncidera pas avec la fonction "idéale" f mais il faut faire en sorte qu'elle n'en soit pas trop éloignée.

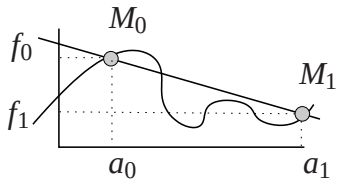
Le problème de l'interpolation polynomiale consiste à choisir comme fonction reconstruite une fonction polynomiale. C'est la méthode la plus ancienne, la plus élémentaire et encore la plus utile. Mais il y en a d'autres. Dans la figure ci-dessus la fonction reconstruite f_r est obtenue à partir de quatre données par un procédé voisin (spline d'interpolation) mais différent.

D'une manière précise, étant donnés $d+1$ points d'abscisses distinctes $M_j = (a_j, f_j)$ $(j = 0, \ldots, d)$ dans le plan (pour des raisons de commodité d'écriture les points seront toujours indicés à partir de 0), le problème consiste à trouver un polynôme $p \in \mathscr{P}_m$ dont le graphe passe par les $d+1$ points M_j. En formule, on doit avoir

$$p \in \mathscr{P}_m \quad \text{et} \quad p(a_j) = f_j \qquad j = 0, \ldots, d \tag{1.5}$$

Ce problème est bien facile à résoudre lorsque lorsque on dispose de deux points M_0 et M_1 et cherche un polynôme de degré 1 car il suffit alors de choisir l'unique polynôme dont le graphe est la droite (M_0M_1) comme indiqué sur la figure.

En effet, posant $p(x) = \alpha x + \beta$, on détermine α et β grâce aux équations $p(a_0) = f_0$ et $p(a_1) = f_1$. On trouve



$$p(x) = \frac{f_1 - f_0}{a_1 - a_0}(x - a_0) + f(a_0) \tag{1.6}$$

que l'on peut aussi écrire

$$p(x) = f(a_0)\frac{x - a_1}{a_0 - a_1} + f(a_1)\frac{x - a_0}{a_1 - a_0}. \tag{1.7}$$

Il est à peine plus compliqué lorsqu'on dispse de trois points $M_i(a_i, f_i)$, $i = 0,1,2$ avec $a_0 < a_1 < a_2$ et cherche un polynôme du second degré. Le graphe cherché est en général une parabole (correspondant à un polynôme de degré 2). Cependant dans le cas particulier où les trois points sont alignés le graphe est à nouveau une droite (correspondant à un polynôme de degré 1).

Ceci dit, s'il n'est pas davantage précisé, le problème (1.5) peut n'avoir aucune solution ou bien en avoir une infinité.

E. 3. (a) Montrer qu'il existe une infinité de polynômes $p \in \mathscr{P}_2$ dont le graphe passe par les points $M_0(0,0)$ et $M_1(1,1)$. (b) Trouver quatre points M_i $(i = 1,2,3,4)$ d'abscisses respectives $-1,0,1,2$ qui ne se trouvent sur le graphe d'aucun polynôme de $\mathscr{P}_2$.

b) Détermination du polynôme d'interpolation

On devine aisément que pour qu'un seul polynôme satisfasse aux conditions (1.5), une relation doit exister entre m et d. Cette relation est facile à mettre en évidence. Pour déterminer $p \in \mathscr{P}_m$, nous devons déterminer l'ensemble de ses coefficients et ceux-ci sont au nombre de $m+1$. Or, pour les déterminer, nous disposons des $d+1$ informations $p(a_i) = f_i$, $i = 0,\ldots,d$. On voit que pour espérer une solution unique, il nous faut supposer que $m = d$. Nous allons démontrer que sous cette condition le problème (1.5) admet effectivement une et une seule solution.

Théorème 1. *Soit $A = \{a_0,\ldots,a_d\}$ un ensemble de $d+1$ nombres réels distincts. Quelles que soient les valeurs $f_0, f_1, \ldots, f_d$, il existe un et un seul polynôme $p \in \mathscr{P}_d$ tel que $p(a_i) = f_i$, $i = 0,1,\ldots,d$. Ce polynôme, est donné par la formule*

$$p = \sum_{i=0}^{d} f_i \ell_i, \tag{1.8}$$

avec

$$\ell_i(x) = \frac{(x-a_0)\cdots(x-a_{i-1})(x-a_{i+1})\cdots(x-a_d)}{(a_i-a_0)\cdots(a_i-a_{i-1})(a_i-a_{i+1})\cdots(a_i-a_d)} \tag{1.9}$$

c) Terminologie et notations

Les nombres a_i s'appellent les **points d'interpolations** ou encore **noeuds d'interpolations**.Lorsque $f_i = f(a_i)$, la fonction f est la **fonction interpolée**. On dit aussi que les valeurs $f(a_i)$ sont les **valeurs interpolées** L'unique polynôme $p \in \mathscr{P}_d$ vérifiant $p(a_i) = f(a_i)$ $(i = 0,1,\ldots,d)$ s'appelle alors le **polynôme d'interpolation de Lagrange** de f aux points a_i. Il est noté $\mathbf{L}[a_0,\ldots,a_d; f]$ ou bien $\mathbf{L}[A; f]$.

Cette dernière notation est parfaitement valable car le polynôme d'interpolation de Lagrange dépend uniquement de l'ensemble des points et non de la manière dont les points sont ordonnés. Une manière un peu sophistiquée de traduire cette propriété est la suivante : si σ est une permutation* quelconque des indices $0,1,\ldots,d$ alors $\mathbf{L}[a_0,\ldots,a_d; f] = \mathbf{L}[a_{\sigma(0)},\ldots,a_{\sigma(d)}; f]$.

Les polynômes ℓ_i s'appellent les **polynômes fondamentaux de Lagrange**. En utilisant le symbole $\prod$ qui est l'équivalent pour le produit de ce que $\sum$ est

* Une permutation des indices $0,1,\ldots,d$ est une bijection de l'ensemble$\{0,1,\ldots,d\}$ dans lui-même.

pour la somme, on a la formule suivante qui est une variante compacte de (1.9).

$$\ell_i(x) = \prod_{j=0, j\neq i}^{d} \frac{x-a_j}{a_i-a_j}. \tag{1.10}$$

Avec ces nouvelles notations, l'expression (1.8) devient

$$\mathbf{L}[a_0,\ldots,a_d; f](x) = \sum_{i=0}^{d} f(a_i) \prod_{j=0, j\neq i}^{d} \frac{x-a_j}{a_i-a_j}. \tag{1.11}$$

Cette expression de $\mathbf{L}[A; f]$ est connue sous le nom de **formule d'interpolation de Lagrange**.

d) Propriétés algébriques et linéarité

Il est essentiel de retenir l'équivalence suivante

$$\left.\begin{array}{c} p \in \mathscr{P}_d \\ p(a_i) = f(a_i) \quad i = 0,\ldots,d \end{array}\right\} \Leftrightarrow p = \mathbf{L}[a_0,\ldots,a_d; f] \tag{1.12}$$

En particulier,

si $p \in \mathscr{P}_d$ alors $\mathbf{L}[a_0,\ldots,a_d; p] = p$.

Il faut prendre garde que cette propriété n'est valable que lorsque le degré de p est inférieur où égal à d.

Cette relation implique des propriétés algébriques intéressantes sur les polynômes ℓ_i. Par exemple, en utilisant que, quel que soit le nombre de points, le polynôme constant égal à 1 est son propre polynome d'interpolation on a

$$\sum_{i=0}^{d} \ell_i = 1. \tag{1.13}$$

E. 4. Vérifiez la propriété ci-dessus par le calcul dans le cas où $d = 1$ (deux points d'interpolation) et $d = 2$ (trois points d'interpolation).

Théorème 2. *l'application qui à f définie (au moins) sur $A = \{a_0,\ldots,a_d\}$ fait correspondre $\mathbf{L}[A; f] \in \mathscr{P}$ est une application linéaire cela signifie qu'elle satisfait les deux propriétés suivantes*

$$\left\{\begin{array}{rcl} \mathbf{L}[A; f+g] & = & \mathbf{L}[A; f] + \mathbf{L}[A; g] \\ \mathbf{L}[A; \lambda f] & = & \lambda \mathbf{L}[A; f] \end{array}\right. . \tag{1.14}$$

E. 5. Montrer les propriétés (1.14).

E. 6. Soit pour tout $n \in \mathbb{N}$, $M_n(x) = x^n$. Déterminer $\mathbf{L}[-1,0,1; M_n]$ et en déduire, pour tout polynôme p, une formule pour $\mathbf{L}[-1,0,1; p]$ en fonction des coefficients de p.

1.3 Algorithme de calcul et exemples graphiques

a) Algorithme basé sur la Formule d'interpolation de Lagrange

Algorithme 3. *Les données de l'algorithme sont*

- *le vecteur* $a = (a_0, \ldots, d)$ *formé des points d'interpolation,*
- *le vecteur* $f = (f_0, \ldots, f_d)$ *formé des valeurs d'interpolations*
- *le point* t *en lequel on veut calculer* $\mathbf{L}[a; f]$.

Le résultat est dans P.

(i) $P := 0$

(ii) Pour $i \in [0 : d]$ *faire*

(a) $L := 1$

(b) Pour $j \in [0 : i-1; i+1 : d]$, $L := L \times (t - a_j)/(a_i - a_j)$

(c) $P := P + L \times f_i$.

b) Exemples

Sur les graphiques de la table 1 on compare la fonction $f(x) = x\sin(\pi x)$ (tracée en bleu) et ses polynômes d'interpolation (tracés en rouge) de degré d par rapport aux $d+1$ **points équidistants** $a_i = -1 + 2i/d$, $i = 0,1,\ldots,d$ lorsque $d = 3$, 4, 5 et 6. Par exemple lorsque $d = 4$, les 5 noeuds d'interpolation sont -1, -0.6, -0.2, 0.2, 0.6, 1. On remarque que les polynômes approchent très bien la fonction de telle manière que les graphes sont confondus sur $[-1,1]$ dès que $d = 6$. Par contre, le résultat est mauvais en dehors de l'intervalle $[-1,1]$. En réalité, avec la fonction choisie, qui est très régulière, en augmentant d on obtiendrait aussi une excellente approximation en dehors de l'intervalle. Nous verrons plus loin des exemples de fonctions pour lesquelles les polynômes d'interpolations construits aux points équidistants ne fournissent pas une bonne approximation. On remarquera que, dans le cas $d = 3$, le graphe du polynôme d'interpolation est une parabole, c'est à dire le graphe d'un polynôme du second degré.

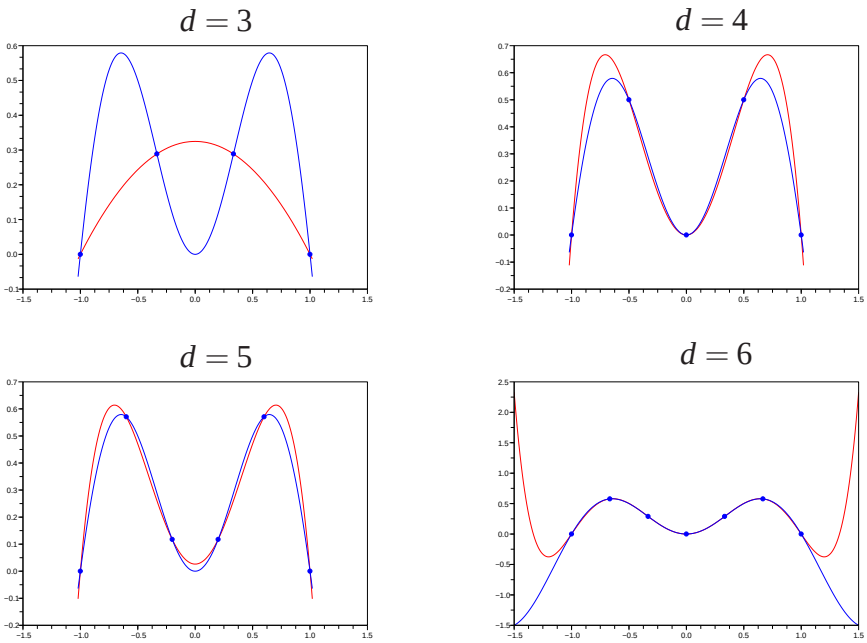
c) La complexité d'un algorithme

(i) Le nombre d'opérations.

(ii) Le type d'opération.

(iii) Le type de nombre.

Théorème 4. *L'algorithme 3 nécessite* $(2d+1)(d+1) \approx 2d^2$ *multiplications-divisions.*



TAB. 1 – *Quelques polynômes d'interpolation de la fonction* $f(x) = x \sin \pi x$.

Ici, le symbole $\approx$ a la sens d'équivalent suivant. On dit que deux suites u_d et v_d sont équivalentes (lorsque $d \to \infty$) et on note $u_d \approx v_d$ lorsque* $\lim_{d\to\infty} u_d/v_d = 1$. Ici, prenant $u_d = (2d+1)(d+1) = 2d^2+3d+1$ et $v_d = 2d^2$ on a $u_d/v_d = 1+3/(2d)+2/(2d^2) \to 0$ lorsque $d \to \infty$.

d) La stabilité d'un algorithme

(i) Prenons 7 points équidistants dans $[-1,1]$, $a_i = -1+2/6 \cdot i$, $i = 0\ldots,6$ et calculons l'interpolant de Lagrange de la fonction $[]^3 : x \to x^3$. D'après on a $\mathbf{L}[_0,\ldots,a_6;[]^3](x) = x^3$. L'algorithme ci-dessus, correctement modifié pour donner un polynôme fournit le résultat donné dans la table 2.

n	coef. de x^n	n	coef. de x^n	n	coef. de x^n
1	$-2.776D-17$	3	1	5	$6.661D-16$
2	$-2.776D-16$	4	$5.551D-16$	6	$-1.110D-15$

TAB. 2

Pour $d = 30$ et la fonction polynôme $p(x) = 6x^2+2x^3+x^4+x^5$, on obtient les coefficients donnés dans table 3.

* La définition donnée ici suppose que la suite v_d ne s'annule jamais, ou au moins ne s'annule pas pour d assez grand. Cette restriction n'est pas nécessaire, dans le cas général, on dit que u_d et v_d sont équivalentes lorsque $u_d = v_d(1+\varepsilon_d)$ ou ε_d est une suite qui tend vers 0.

n	coef. de x^n	n	coef. de x^n	n	coef. de x^n
1	$-7.154D-16$	11	0.0000102	21	0.0312387
2	6	12	-0.0000084	22	-0.0128725
3	2	13	0.0000207	23	-0.0083289
4	1	14	-0.0003492	24	0.0210372
5	1	15	0.0021007	25	-0.0186058
6	$2.804D-09$	16	-0.0073588	26	0.0090390
7	$3.535D-09$	17	0.0175530	27	-0.0024166
8	0.0000001	18	-0.0304909	28	0.0003477
9	0.0000011	19	0.0399997	29	0.0000052
10	-0.0000046	20	-0.0407562	30	-0.0000115

TAB. 3 – *Coefficients de l'interpolant de Lagrange de* $p(x) = 6x^2+2x^3+x^4+x^5$ *avec* 30 *points équidistants dans* $[-1,1)$.

(ii) Explication du résultat inexact.

(iii) Définition informelle de l'instabilité. La stabilité dépend de :

(a) Les points d'interpolation $a_0 \ldots, a_d$. De ce points de vue les points équidistants constituent un mauvais choix.

(b) La fonction interpolée. Les risques d'erreur sont importants si la fonctions admet des variations importantes c'est-à-dire lorsque $f(x+\varepsilon)$ peut être très différent de $f(x)$ pour ε petit. C'est le cas des fonctions avec $f'(x)$ grand. Illustration.

(c) L'algorithme utilisé, dont les qualités dépendent de la méthode mathématique dont il découle

– du programme ou langage à l'intérieur duquel l'algorithme est programmé,

– de l'habilité du traducteur,

1.4 L'algorithme de Neville-Aitken

a) La formule de récurrence de Neville-Aitken

Théorème 5 (Neville-Aitken)**.** *Soit* $A = \{a_0, a_1, \ldots, a_d\}$ *un ensemble de* $d+1$ *nombres réels distincts et* f *une fonction définie (au moins) sur* A*. On a*

$$(a_0 - a_d)\mathbf{L}[a_0, a_1, \ldots, a_d; f](x) = (x - a_d)\mathbf{L}[a_0, a_1, \ldots, a_{d-1}; f](x) - (x - a_0)\mathbf{L}[a_1, a_2, \ldots, a_d; f](x). \quad (1.15)$$

Corollaire. *Sous les mêmes hypothèses, pour tout couple d'indice* (i,j) *dans* $\{0,\dots,d\}$ *avec* $i \neq j$*, on a*

$$(a_i - a_j)\mathbf{L}[a_0,a_1,\dots,a_d;f](x) = (x-a_j)\mathbf{L}[a_0,\dots,a_{j-1},a_{j+1},\dots a_d;f](x) - (x-a_i)\mathbf{L}[a_0,\dots,a_{i-1},a_{i+1},\dots a_d;f](x). \quad (1.16)$$

b) L'algorithme

On pose $A = \{x_1,x_2,\dots,x_{d+1}\}$ un ensemble de $d+1$ réels distincts. On notera que les points sont indicés à partir de 1 et non pas comme jusqu'à présent à partir de 0. On définit une famille (triangulaire) de polynômes $p_{i,m}$ par récurrence sur $m \in \{0,1,\dots,d\}$ comme suit

$$p_{i,0}(x) = f(x_i), \quad 1 \leq i \leq d+1 \quad (1.17)$$

puis,

$$p_{i,m+1}(x) \quad = \frac{(x_i - x)p_{m+1,m}(x) - (x_{m+1} - x)p_{i,m}(x)}{x_i - x_{m+1}}, \quad m+2 \leq i \leq d+1. \quad (1.18)$$

Théorème 6. *Pour* $0 \leq m \leq d$*, on a*

$$p_{i,m} = \mathbf{L}[x_1,x_2,\dots,x_m,x_i;f] \quad (m+1 \leq i \leq d+1). \quad (1.19)$$

Lorsque $m = 0$ *l'écriture* $\mathbf{L}[x_1,x_2,\dots,x_m,x_i;f]$ *doit être comprise comme* $\mathbf{L}[x_i;f]$*. En particulier*

$$\mathbf{L}[x_1,\dots,x_{d+1};f] = p_{d+1,d}.$$

Algorithme 7. *Les données de l'algorithme sont*

- *le vecteur* $x = (x_1,\dots,x_{d+1})$ *formé des points d'interpolation,*
- *le vecteur* $f = (f_1,\dots,f_{d+1})$ *formé des valeurs d'interpolations*
- *le point* t *en lequel on veut calculer* $\mathbf{L}[x;f]$*.*

On utilise une matrice P *de dimension* $(d+1)\times(d+1)$ *que l'on initialise à* 0*. Le résultat est dans* $P(d+1,d+1)$*.*

(i) Pour $j \in [1:d+1]$*,* $P(j,1) = y(j)$*.*

(ii) Pour $m \in [2:d+1]$

Pour $i \in [m:d+1]$

$$p(i,m) = \frac{\big(x(i)-x\big)\times p(m-1,m-1) - \big(x(m-1)-x\big)\times p(i,m-1)}{(x(i)-x(m-1))} \quad (1.20)$$

La table 4 reprend l'exemple de la table 3 précédent et donne les six plus mauvais coefficients obtenus en utilisant l'algorithme de Neville ci-dessus.

n	coef. de x^n	n	coef. de x^n	n	coef. de x^n
17	0.0024102	21	−0.0020416	24	−0.0019227
18	0.0023906	22	0.0026459	26	0.0022589

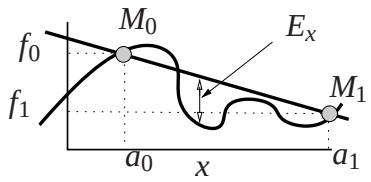
TAB. 4

§ 2 ETUDE DE L'ERREUR

2.1 L'énoncé du théorème

Comme le polynôme d'interpolation $\mathbf{L}[a_0,\ldots,a_d\,;\,f]$ est égal à la fonction f en tous les points a_i, $i=0,\ldots,d$, il est naturel d'espérer que la différence entre f et ce polynôme aux autres points sera petite c'est-à-dire, que $\mathbf{L}[a_0,\ldots,a_d\,;\,f]$ fournira une bonne approximation de $f(x)$, au moins en les points x pas trop éloignés des a_i.

Pour mesurer la qualité de cette approximation, nous devrons estimer (majorer) l'erreur E_x entre $f(x)$ et $\mathbf{L}[a_0,\ldots,a_d;\,f](x)$. La figure ci-contre fait apparaître cette erreur dans le cas $d=1$. Cette erreur est une distance,



$$E_x = |f(x) - \mathbf{L}[a_0,\ldots,a_d;\,f](x)|.$$

On devine facilement qu'elle dépendra à la fois de la fonction f et de la position des points a_i. Le théorème suivant, et surtout son corollaire, fournissent une estimation simple de l'erreur.

Théorème 8. *Soient $f \in \mathrm{C}^{d+1}[a,b]$ et $A=\{a_0,a_1,\ldots,a_d\}\subset[a,b]$. Nous supposons que $a_0<a_1<a_2<\cdots<a_{d-1}<a_d$. Pour tout $x\in[a,b]$, il existe $\xi=\xi_x\in\,]a,b[$ tel que*

$$f(x)-\mathbf{L}[A;f](x)=\frac{f^{(d+1)}(\xi)}{(d+1)!}(x-a_0)(x-a_1)\ldots(x-a_d). \qquad (2.1)$$

Rappelons que $f\in\mathrm{C}^{d+1}[a,b]$ signifie que f est $d+1$ fois dérivable et que $f^{(d+1)}$, la dérivée $(d+1)$-ième est continue. Au point a (resp. b) il s'agit de dérivées à droite (resp. à gauche).

Lorsque $d=0$ et $A=\{a_0\}$, on a $\mathbf{L}[a_0\,;\,f](x)=f(a_0)$ de sorte que le théorème 8 affirme que, pour un x fixé dans [a,b], il existe un point ξ – dépendant de x – tel que

$$f(x)-f(a_0)=f'(\xi)(x-a_0).$$

Il s'agit de la *formule des accroissements finis* dont le théorème 8 est par conséquent une généralisation.

E. 7. Soit $a \leq a_0 < a_1 \leq b$. Montrer que si f est une fonction strictement convexe deux fois dérivable sur $[a,b]$ alors $f(x) - \mathbf{L}[a_0,a_1\,;\,f](x) < 0$ pour tout $x \in]a_0,a_1[$. Que dire en dehors de l'intervalle $[a_0,a_1]$? Même question dans le cas des fonctions deux fois dérivables strictement concaves. Étudier le problème sans supposer que les fonctions soient deux fois dérivables.

Dans la pratique, le corollaire suivant est très souvent suffisant.

Corollaire.

$$|f(x) - \mathbf{L}[A;f](x)| \leq \frac{\|f^{(d+1)}\|_\infty}{(d+1)!}|x-a_0||x-a_1|\dots|x-a_d| \tag{2.2}$$

où $\|f^{(d+1)}\|_\infty = \sup_{x\in[a,b]}|f^{(d+1)}(x)|$. *En particulier,*

$$\|f - \mathbf{L}[A;f]\|_\infty \leq \frac{\|f^{(d+1)}\|_\infty}{(d+1)!}\|w_A\|_\infty, \tag{2.3}$$

où w_A *est le polynôme de degré* $d+1$ *défini par*

$$w_A(x) = (x-a_0)(x-a_1)\cdots(x-a_d). \tag{2.4}$$

E. 8. Considérons les réels $a_0 = 100$, $a_1 = 121$ et $a_2 = 144$ et la fonction f définie de $\mathbb{R}^+$ dans lui-même par $f(x) = \sqrt{x}$. Calculer $\mathbf{L}[a_0,a_1,a_2;f](115)$ et montrer que

$$\left|\sqrt{115} - \mathbf{L}[a_0,a_1,a_2;f](115)\right| < 1{,}8 \cdot 10^{-3}.$$

La démonstration du théorème 8, assez délicate, sera donnée un peu plus loin après que nous nous serons munis de l'outil nécessaire qui est un théorème de Rolle généralisé.

2.2 Le théorème de Rolle généralisé

Rappelons que théorème de Rolle ordinaire affirme que si f est une fonction continue sur $[a,b]$ et dérivable sur $]a,b[$ telle que $f(a) = f(b) = 0$ alors il existe c tel que $f'(c) = 0$. Ici, nous aurons besoin du

Théorème 9 (de Rolle généralisé)**.** *Si* u *est un fonction continue sur* $[a,b]$ *et* k *fois dérivable sur* $]a,b[$ *qui admet* $k+1$ *zéros* x_i, $i = 0,\dots,k$, *alors il existe* $c \in]a,b[$ *tel que* $u^{(k)}(c) = 0$.

L'énoncé habituel du théorème de Rolle correspond au cas $k = 1$.

E. 9. Donner une démonstration par récurrence du théorème.

2.3 Démonstration du théorème 8

2.4 Conséquence de la formule d'erreur sur le choix des points d'interpolation

Le second corollaire du theorème 8 montre que si vous nous voulons rendre l'erreur entre la fonction et son polynôme interpolation la plus petite possible et que nous sommes libres de choisir les points d'interpolation a_i, $i = 0,\ldots,a_d$, comme nous le voulons dans $[a,b]$, alors nous avons intérêt à choisir ces points de telle sorte que la quantité $\|w_A\|_\infty$, voir (2.4), soit la plus petite possible. Il existe un unique ensemble de points qui minimise cette quantité. On les appelle les **points de Chebyshev** en hommage au mathématicien russe qui les a déterminés pour la première fois en 1874. Lorsque $[a,b] = [-1,1]$ ces points sont donnés par la formule

$$a_i = \cos\left(\frac{2i+1}{2(d+1)}\,\pi\right), \quad i = 0,\ldots,d. \tag{2.5}$$

La figure 1 compare la répartition des points de Chebyshev et des points équidistants, donnés lorsque $[a,b] = [-1,1]$ par la formule $a_i = 1+2i/d$, lorsque $d = 50$. On remarquera que les premiers tendent à se densifier lorsque qu'on approche des extrémités de l'intervalle.

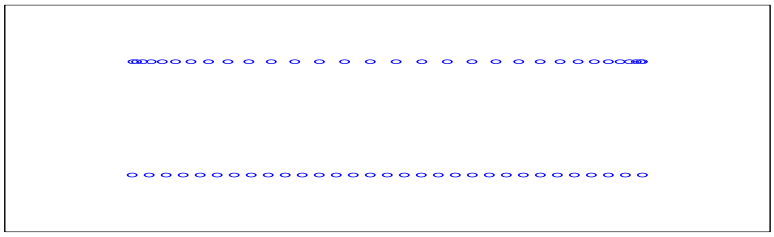
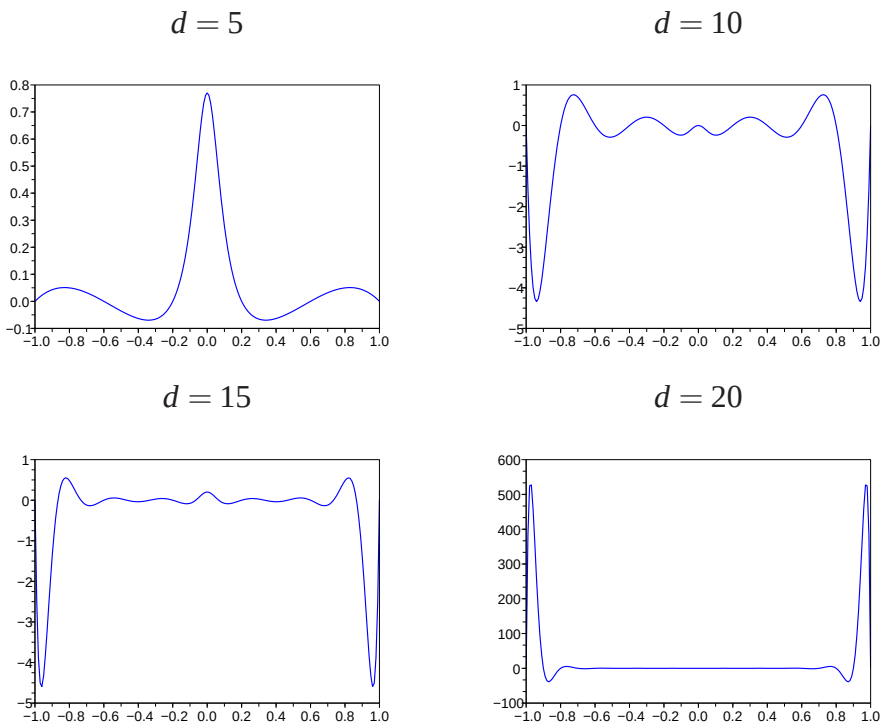


FIG. 1 – *Répartition des points de chebyshev et des points equidistants ($d = 30$)*

2.5 Précision de l'interpolant et nombre de points d'interpolation

Il est naturel de penser que plus on augmente le nombre de points d'interpolation, meilleure est la précision de l'approximation fournie par le polynôme d'interpolation. Cette intuition est renforcée par les exemples donnés dans la table 1. Pourtant, si cette idée reste correcte pour une classe importante de fonctions et pour des points d'interpolation correctement choisis, elle est fausse dans le cas général. En particulier, quels que soient les points d'interpolation choisis, il est toujours possible de trouver une fonction continue qui ne se laisse pas approcher

par les polynômes interpolation correspondant. L'exemple le plus classique a été donné par le mathématicien Runge qui a montré en 1901 que les polynômes d'interpolation aux points équidistants de la fonction f définit par $f(x) = 1/(1+x^2)$ donnaient des résultats très mauvais. La table 2.5 donne le graphe de la fonction d'erreur entre le polynôme interpolation au point équidistant et la fonction de Runge modifiée $f(x) = 1/(1+100x^2)$ pour quelques valeurs de d. Ici nous avons modifié la fonction de Runge classique pour accélérer le phénomène de divergence.

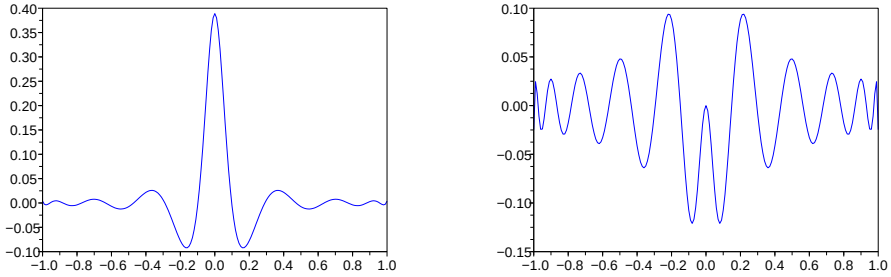


TAB. 5 – *Graphe de la fonction d'erreur entre la fonction* $f(x) = \frac{1}{1+100x^2}$ *et ses interpolants de Lagrange aux points équidistants lorsque* $d = 5,10,15$ *et* 20

Par contre, il est possible de montrer que les polynômes d'interpolation aux points de Chebyshev convergent* vers la fonction interpolée, lorsque le nombre de points croît indéfiniment, sous la seule condition que la fonction soit dérivable, de dérivée bornée. La convergence cependant peut être lente. La table 2.5 reprend l'exemple de la fonction de Runge et donne la fonction d'erreur entre cette fonction et le polynôme d'interpolation aux points de Chebyshev.

∗ Il s'agit ici de convergence uniforme des fonctions. Cela signifie que la suite de nombres réels positifs $\|f - \mathbf{L}[a_0, \cdots, a_d\,; f]\|_\infty$, $d \in \mathbb{N}$, converge vers 0 lorsque $d \to \infty$.

$d = 15$ $d = 20$



TAB. 6 – *Graphe de la fonction d'erreur entre la fonction* $f(x) = \frac{1}{1+100x^2}$ *et ses interpolants de Lagrange aux points de Chebyshev lorsque* $d = 15$ *et* $d = 20$

§ 3 FONCTIONS POLYNOMIALES PAR MORCEAUX

3.1 Introduction

a) Subdivisions

On appelle **subdivision de longueur** d de $I = [a,b]$ une suite (strictement) croissante de $d+1$ éléments de I, $\sigma = (a_0, \cdots, a_d)$ telle que $a_0 = a$ et $a_d = b$. Autrement dit

$$a = a_0 < a_1 < a_2 < \cdots < a_{d-1} < a_d = b. \tag{3.1}$$

A chaque subdivision σ de longueur d de $[a,b]$ est associée une **partition** de l'intervalle $[a,b]$,

$$[a,b] = [a_0\,,a_1[\cup[a_1\,,a_2[\cup\ldots[a_{d-2}\,,a_{d_1}[\cup[a_{d-1}\,,a_d]. \tag{3.2}$$

La distance entre deux points successif a_i et a_{i+1} est noté h_i et l'**écart** h de la subdivision σ est la plus grande des distances entre deux points successifs,

$$h = \max_{i=0,\ldots,d} h_i = \max_{i=0,\ldots,d-1} (a_{i+1} - a_i). \tag{3.3}$$

Ces définitions sont mises en évidence sur la figure 2.

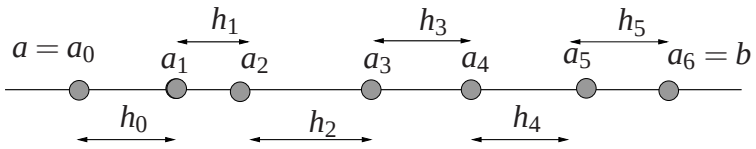


FIG. 2

Lorsque les distances h_i sont constantes, $h_i = (b-a)/d$, la subdivision est formée des points équidistants

$$\sigma = \left(a + i\frac{b-a}{d} \,:\, i = 0, \ldots, d \right).$$

On dit qu'une fonction g est **affine par morceaux** sur l'intervalle I s'il existe une subdivision $\sigma = (a_0, \ldots, a_d)$ de l'intervalle I telle que la restriction de g à chacun des sous-intervalles défini par σ est une fonction affine, c'est-à-dire pour $i = 0, \ldots, d-1$, il existe des coefficients α_i et β_i tels que

$$x \in [a_i\,, a_{i+1}[\implies g(x) = \alpha_i x + \beta_i.$$

Remarquons que cette définition n'impose aucune condition sur la valeur de g à l'extrémité $b = a_d$ de l'intervalle.

E. 10. A quelles conditions (sur les nombres α_i et β_i) la fonction g est-elle continue? (continue et) convexe? Que dire de la dérivabilité des fonctions affines par morceau?

b) Fonctions polylignes

Soit σ une subdivision de $[a,b]$ et $f = (f_0, \ldots, f_d)$ un ensemble de valeurs quelconques. Nous pouvons construire les polynômes de Lagrange $\mathbf{L}[a_i\,, a_{i+1}\,;\, f_i\,, f_{i+1}]$ pour $i = 0, \ldots, d-1$, c'est-à-dire les uniques polynômes de degré inférieur ou égal à 1 qui prennent les valeurs f_i au point a_i et f_{i+1} au point a_{i+1}. La fonction **polyligne** associée à la subdivision σ et aux valeurs f, notée $\mathbf{PL}[\sigma, f]$, est définie sur chacun des sous-intervalles définis par la subdivision par la relation

$$\begin{cases} \mathbf{PL}[\sigma\,, f](x) = \mathbf{L}[a_i\,, a_{i+1}\,;\, f_i\,, f_{i+1}](x), & x \in [a_i\,, a_{i+1}[\\ \mathbf{PL}[\sigma\,, f](b) = f_d \end{cases} \tag{3.4}$$

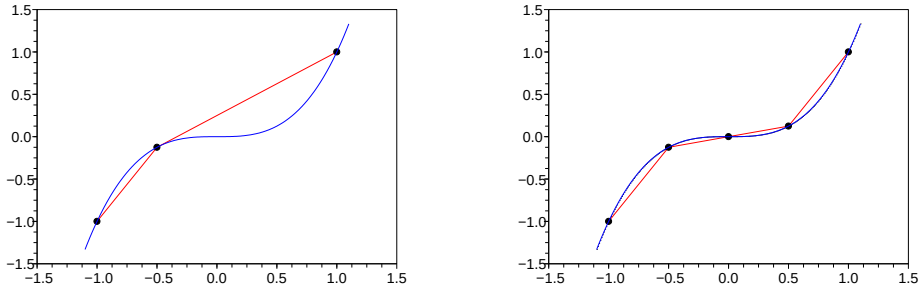
Lorsque les valeurs f_i sont les valeurs d'une fonction f aux points a_i, $f_i = f(a_i)$, $i = 0\,, \ldots\,, d$, on dit que $\mathbf{PL}[\sigma; f]$ est la (fonction) **polyligne** interpolant la fonction f aux points de la subdivision σ.

Les deux schémas dans le tableau 7 font apparaître en rouge les graphes des fonctions $\mathbf{PL}[\sigma; f]$ lorsque $f(x) = x^3$ (tracé en bleu) et (1) $\sigma = (-1, -0.5, 1)$ puis (2) $\sigma = (-1, -0.5, 0, 0.5, 1)$.

Théorème 10. $\mathbf{PL}[\sigma, f]$ *est une fonction affine par morceaux continue satisfaisant*

$$\mathbf{PL}[\sigma, f](a_i) = f_i, \quad i = 0, \ldots, d. \tag{3.5}$$

E. 11. Expliquer pourquoi la fonction $\mathbf{PL}[\sigma, f]$ n'est pas la seule fonction affine par morceaux vérifiant les conditions du théorème. Quelle propriété supplémentaire, non formulée dans le théorème, caractérise-t-elle $\mathbf{PL}[\sigma, f]$?



(1) (2)

TAB. 7 – *Deux exemples de polylignes*

c) Approximation par les fonctions polylignes

Au contraire des polynômes d'interpolation de Lagrange, les fonctions polylignes fournissent une bonne approximation de toutes les fonctions continues, pour peu que l'écart de la subdivision soit suffisamment petit. Ici nous nous contentons d'énoncer et de démontrer un théorème qui concerne les fonctions dérivables, de dérivés continues.

Théorème 11. *Soit f une fonction continûment dérivable sur $[a,b]$ et σ une subdivision de $[a,b]$ d'écart h. Pour tout fixes $x \in [a,b]$, on a*

$$|f(x) - \mathbf{PL}[\sigma; f](x)| \leq h \cdot \max_{t\in[a,b]} |f'(t)|. \tag{3.6}$$

Corollaire. *Si σ^d, $d \in \mathbb{N}$, est une suite de subdivisions de longueur d de $[a,b]$ dont l'écart tend vers 0 lorsque d tend vers ∞ on a*

$$\lim_{d\to\infty} \mathbf{PL}[\sigma^d; f] = f(x), \quad x \in [a,b]. \tag{3.7}$$

S'agissant d'une suite de subdivisions, à chaque changement de d, les points de la subdivision changent, excepté le premier qui doit toujours être égal à a et le second qui doit toujours être égal à b,

$$\sigma^d = (a, a_1^d, a_2^d, \ldots, a_{d-1}^d, b).$$

naturellement, l'écart de la subdivision σ^d dépend de d.

Corollaire. *Lorsque σ^d est la subdivision formée des points équidistants*

$$a_i = a + i \cdot \frac{b-a}{d}, \quad i = 0, \ldots, d+1, \quad d \in \mathbb{N}^\star$$

alors

$$|f(x) - \mathbf{PL}[\sigma; f](x)| \leq \frac{b-a}{d} \cdot \max_{t\in[a,b]} |f'(t)| \overset{d\to\infty}{\to} 0, \quad x \in [a,b]. \tag{3.8}$$

3.2 Représentation

Nous allons déterminer des fonctions $b_i = b_i^\sigma$ adaptées à la subdivision σ qui permettent une représentation simple du polynôme $\mathbf{PL}\sigma; f$. Pour que tous les points a_i, $i = 0,\ldots,d$ jouent un rôle semblable on est amené à compléter la subdivision σ par deux points a_{-1} et a_{d+1} comme indiqué sur la figure 3. Ces points peuvent être choisis librement sous les seules conditions que $a_1 < a = a_0$ et $a_{d+1} > a_d = b$.

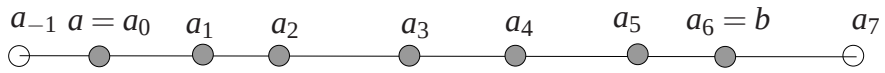


FIG. 3 – *Sudvision complétée des points a_{-1} et a_{d+1}.*

Une fois la subdivision complétée, nous définissons pour $i = 0,\ldots,d$ la fonction b_i sur $\mathbb{R}$ par le graphe donné dans la figure 3.2.

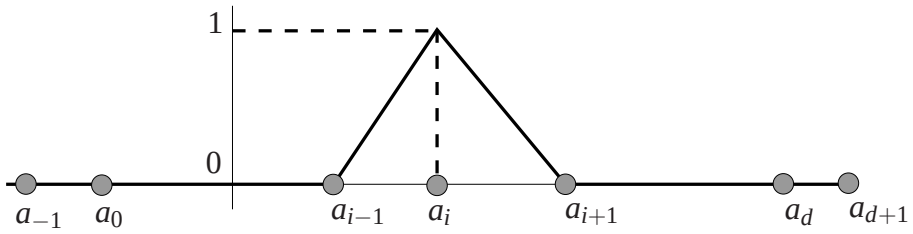


FIG. 4 – *graphe de la fonction b_i.*

En formule, la fonction b_i est définie par

$$b_i(x) = \begin{cases} 0 & \text{if } x \le a_{i-1} \\ \frac{x-a_{i-1}}{a_i-a_{i-1}} & \text{if } a_{i-1} \le x \le a_i \\ \frac{x-a_{i+1}}{a_i-a_{i+1}} & \text{if } a_i \le x \le a_{i+1} \\ 0 & \text{if } x \ge a_{i+1}. \end{cases} \tag{3.9}$$

Les fonctions b_i, $i = 0,\ldots,d$, sont affines par morceaux, continues et s'annulent en tout les points a_j sauf lorsque $j = i$ auquel cas on a $b_i(a_i) = 1$. Remarquons aussi qu'elles sont nulles en dehors de l'intervalle $[a_{i-1}, a_{i+1}$. Cet intervalle est appelé le **support** de la fonction b_i.

Théorème 12. $\mathbf{PL}[\sigma; f](x) = \sum_{i=0}^{d} f_i b_i(x)$.

3.3 Extension

II

CALCUL APPROCHÉ DES INTÉGRALES

§ 1 FORMULES DE QUADRATURES ÉLÉMENTAIRES

1.1 Problème

Soit $f \in \mathsf{C}[a,b]$. On souhaite calculer $\int_a^b f(x)dx$. Le théorème fondamental du calcul intégral nous dit que $\int_a^b f(x)dx = F(b) - F(a)$ où F est une primitive de f. Pour déterminer une primitive on dispose de quelques outils théoriques dont les plus élémentaires sont le théorème d'intégration par partie et le théorème de changement de variable. Mais on est capable de déterminer explicitement F pour une classe relativement restreinte de fonctions f. Même lorsqu'il est possible de déterminer F, son expression est souvent si compliquée que l'évalutaion de $F(b) - F(a)$ nécessite l'emploi d'un processus d'approximation et, dans ce cas, il est plus naturel et généralement moins coûteux de chercher directement une approximation de l'intégrale.

1.2 Présentation générale

L'idée consiste à utiliser une approximation $\int_a^b f(x)dx \approx \int_a^b g(x)dx$ où g est est une fonction qui d'une part est proche de f et d'autre part possède des primitives facilement calculables. Le choix le plus naturel est $g = \mathbf{L}[x_0,\ldots,x_d; f]$ où $A = \{x_0,\ldots,x_d\} \subset [a,b]$ car les polynômes d'interpolation de Lagrange sont proches de la fonction qu'ils interpolent et, étant des polynômes, on espère que leurs primitives sont facilement calculables. On appelle **formule de quadrature**

(élémentaire) d'ordre d, toute expression

$$\mathtt{Q}(f) = \int_a^b \mathbf{L}[x_0,\ldots,x_d;f](x)dx = \sum_{i=0}^{d} f(x_i)\int_a^b \ell_i(x)dx \tag{1.1}$$

où ℓ_i est le polynôme fondamental de Lagrange correspondant au point a_i, L'application Q ainsi définie est une forme linéaire sur $\mathsf{C}[a,b]$, autrement dit elle vérifie

$$\mathtt{Q}(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1\mathtt{Q}(f_1) + \lambda_2\mathtt{Q}(f_2) \qquad (\lambda_1,\lambda_2 \in \mathbb{R}, f_2, f_2 \in \mathbb{C}[a,b]).$$

Pour savoir si $\mathtt{Q}(f)$ est effectivement proche de $\int_a^b f(x)dx$ on devra estimer l'erreur

$$\mathsf{E}^{\mathtt{Q}}(f) := \left|\int_a^b f(x)dx - \mathtt{Q}(f)\right| \tag{1.2}$$

Remarquons que si $\mathtt{Q}$ est une formule de quadrature d'ordre d alors pour tout $p \in \mathscr{P}_d$ on a $\int_a^b p(x)dx = \mathtt{Q}(p)$. En effet,

$$\begin{aligned} p \in \mathscr{P}_d \quad &\Rightarrow \quad p = \mathbf{L}[x_0,\ldots,x_d;p] \\ &\Rightarrow \quad \mathtt{Q}(p) \overset{def}{=} \int_a^b \mathbf{L}[x_0,\ldots,x_d;p](x)dx = \int_a^b p(x)dx. \end{aligned}$$

Nous verrons que dans certains cas l'égalité ci-dessus peut continuer à être vérifiée pour des polynômes de degré plus grand que d.

Une réciproque est vraie.

Théorème 1. *Si $R(f)$ est une expression de la forme $R(f) = \sum_{i=0}^{d}\lambda_i f(a_i)$ telle que $R(p) = \int_a^b p(x)dx$ pour tout $p \in \mathscr{P}_d$ alors $\lambda_i = \int_a^b \ell_i(x)dx$ où ℓ_i est le polynôme fondamental de Lagrange correspondant à $x_i \in \{x_0,\ldots,x_d\}$.*

E. 12. On cherche une approximation de $\int_{-1}^{1} f(x)dx$ par une formule du type

$$\int_{-1}^{1} f(x)dx \approx f(t_1) + f(t_2)$$

de telle sorte que la formule soit *exacte* pour tous les polynômes de dégré inférieur ou égal à 2. Montrer qu'il existe une et une seule paire $\{t_1,t_2\}$ satisfaisant la propriété demandée et la déterminer.

Dans la pratique, grâce au procédé de composition, on obtient souvent des résultats très précis en employant seulement des méthodes d'ordre $d \leq 2$. Nous étudierons en détail trois de ces méthodes : la *méthode du point milieu* ($d = 0$), la *méthode des trapèzes* ($d = 1$) et la méthode de Simpson ($d = 2$).

1.3 Exemples fondamentaux : formules du point milieu, du trapèze et de Simpson

a) Le formule du point milieu

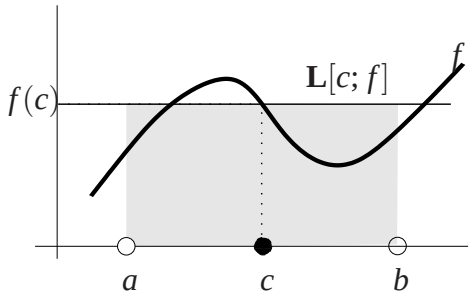
On utilise un polynôme d'interpolation de degré $d = 0$ avec le point $x_0 = \frac{a+b}{2}$. Dans ce cas $\mathbf{L}[x_0; f](x) = f(\frac{a+b}{2})$ et l'approximation

$$\int_a^b f(x)dx \approx \int_a^b \mathbf{L}[x_0; f](x)dx \quad \text{devient} \quad \int_a^b f(x)dx \approx (b-a)\, f(\frac{a+b}{2}). \tag{1.3}$$

L'expression

$$\mathtt{Q}(f) = (b-a)\, f(\frac{a+b}{2}) \tag{1.4}$$

s'appelle la **formule du point milieu**.



On a posé $c = (a+b)/2$. L'aire de la partie grisée est égale à $\mathtt{Q}(f)$.

FIG. 1 – *Méthode du point milieu.*

b) La formule du trapèze

Soit $f \in \mathrm{C}[a,b]$. On prend $d = 1$ et $A = \{a,b\}$. L'approximation

$$\int_a^b f(x)dx \approx \int_a^b \mathbf{L}[a,b; f](x)dx \quad \text{devient} \quad \int_a^b f(x)dx \approx \frac{(b-a)}{2}(f(a)+f(b)). \tag{1.5}$$

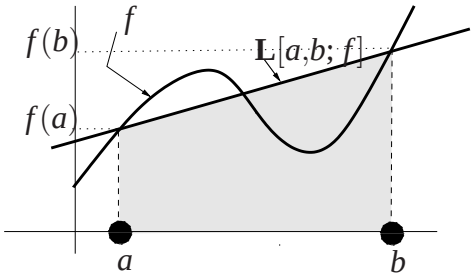
En effet $\mathbf{L}[a,b;f](x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$ d'où

$$\begin{aligned}
\int_a^b \mathbf{L}[a,b;f](x)dx &= \int_a^b f(a) + \left\{ \frac{f(b)-f(a)}{b-a}(x-a) \right\} dx \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a} \int_a^b (x-a)dx \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a} \left[\frac{(x-a)^2}{2} \right]_a^b \\
&= f(a)(b-a) + \frac{f(b)-f(a)}{b-a} . \frac{(b-a)^2}{2} \\
&= \frac{(b-a)}{2} . [f(a)+f(b)].
\end{aligned}$$

L'expression

$$\mathtt{Q}(f) = \frac{(b-a)}{2}(f(a)+f(b)) \tag{1.6}$$

s'appelle **la formule du trapèze**.



Com. L'aire de la partie grisée est égale à $\mathtt{Q}(f)$.

FIG. 2 – *Méthode des trapèzes.*

c) Formule de Simpson

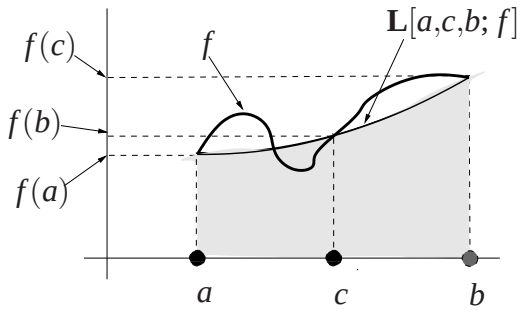
On prend $d = 2$ et $A = \{a,c,b\}$ où $c = \frac{a+b}{2}$. L'approximation $\int_a^b f(x)dx \approx \int_a^b \mathbf{L}[a,c,b;f](x)dx$ devient (cf. exercices)

$$\int_a^b f(x)dx \approx \frac{(b-a)}{6}\left(f(a)+4f(c)+f(b)\right). \tag{1.7}$$

L'expression

$$\mathtt{Q}(f) = \frac{(b-a)}{6}\left(f(a)+4f(c)+f(b)\right)$$

s'appelle **la formule de Simpson**.



Com. L'aire de la partie grisée est égale à Q(f).

FIG. 3 – *Méthode de Simpson.*

§ 2 ETUDE DE L'ERREUR

2.1 Estimation de l'erreur dans la formule du point milieu

Théorème 2. *Soit $f \in C^2([a,b])$. Il existe $\xi \in [a,b]$ tel que*

$$\int_a^b f(t)dt - (b-a)\,f\left(\frac{a+b}{2}\right) = \frac{(b-a)^3}{24}\cdot f^{(2)}(\xi).$$

En particulier,

$$\left|\int_a^b f(t)dt - (b-a)\,f\left(\frac{a+b}{2}\right)\right| \leq \frac{(b-a)^3}{24}\cdot \sup_{[a,b]}|f^{(2)}|.$$

Comme de nombreuses questions d'analyse numérique, la démonstration de ce théorème est basée sur la **formule de Taylor**. Nous la rappelons dans un cadre suffisamment général pour pouvoir servir dans la suite du cours.

Théorème 3 (Formule de Taylor). *Soit f une fonction continue sur $[\alpha,\beta]$ et $d+1$ fois dérivable sur $]\alpha,\beta[$. Si u_0 et v sont dans $[\alpha,\beta]$ alors il existe $\xi \in]\alpha,\beta[$ tel que*

$$f(v) = f(u_0) + f'(u_0)(v-u_0) + \cdots + \frac{f^{(d)}(u_0)}{d!}(v-u_0)^d + \frac{f^{(d+1)}(\xi)}{d!}(v-u_0)^{d+1}. \tag{2.1}$$

Cette égalité s'appelle la formule de Taylor de f en u_0 à l'ordre d.

Dans ce cours nous appliquerons toujours ce théorème avec une fonction f de classe C^{d+1} sur un intervalle contenant α et β de sorte que les conditions du théorème seront largement satisfaites.

2.2 Estimation de l'erreur dans la formule du trapèze

Théorème 4. *Soit $f \in C^2([a,b])$. Il existe $\theta \in [a,b]$ tel que*

$$\int_a^b f(t)dt - \frac{(b-a)}{2}[f(a)+f(b)] = -\frac{(b-a)^3}{12} f^{(2)}(\theta).$$

En particulier,

$$\left| \int_a^b f(t)dt - \frac{(b-a)}{2}[f(a)+f(b)] \right| \leq \frac{(b-a)^3}{12} . \sup_{[a,b]} |f^{(2)}|.$$

2.3 Estimation de l'erreur dans la formule de Simpson

Théorème 5 (‡). *Soit $f \in C^4([a,b])$. On a*

$$\left| \int_a^b f(t)dt - \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \right| \leq \frac{(b-a)^5}{2880} . \sup_{[a,b]} |f^{(4)}|.$$

§ 3 COMPOSITION

3.1 Idée générale

Nous que le polynôme $\mathbf{L}[x_0,\ldots,x_d;f]$ est d'autant plus de chance d'être proche de la fonction interpolée f que l'intervalle $[a,b]$ est petit. Les formules d'erreur précédente confirme l'intuition que plus l'intervalle $[a,b]$ est petit plus l'approximation sera précise. Il est naturel de découper l'intervalle de départ en une famille de sous-intervalle beaucoup plus petits et d'appliquer les formules de quadratures à ces petits intervalles. De manière précise, on choisit une subdivision $\sigma = (a = a_0, a_1, \ldots, a_n = b)$ de $[a,b]$ et, dans chaque intervalle $[a_i,a_{i+1}]$, on choisit $d+1$ points distincts $X^i = \{x_0^i, x_1^i, \ldots, x_d^i\}$ pour construire la formule d'approximation

$$\int_{a_i}^{a_{i+1}} f(x)dx \approx \mathtt{Q}_{[a_i,a_{i+1}]}(f) \tag{3.1}$$

avec

$$\mathtt{Q}_{[a_i,a_{i+1}]}(f) = \sum_{i=0}^{d} f(x_i^k) \int_{a_k}^{a_{k+1}} \ell_i^k(x)dx \quad \text{et} \quad \ell_i^k(x) = \prod_{j=0, j\neq i}^{d} \frac{x - x_j^k}{x_i^k - x_j^k}. \tag{3.2}$$

La relation de Chasle pour les intégrales nous donne $\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{a_k}^{a_{k+1}} f(x)dx$ de sorte que pour approximer l'intégrale globale il suffit d'approximer les n termes de la somme

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n-1} \sum_{i=0}^{d} \mathtt{Q}_{[a_k,a_{k+1}]}(f) \tag{3.3}$$

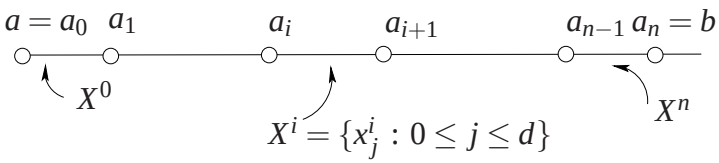


FIG. $4-k$

Toute expression $\mathtt{Qc}$ de la forme

$$\mathtt{Qc}(f) = \sum_{k=0}^{n-1} \sum_{i=0}^{d} \mathtt{Q}_{[a_k,a_{k+1}]}(f)$$

s'appelle une **formule de quadrature composée** d'ordre d. L'application $\mathtt{Qc}$ définit une forme linéaire sur $\mathsf{C}[a,b]$. On note $\mathsf{E}^{\mathtt{Qc}}(f) = |\int_a^b f(x)dx - \mathtt{Qc}(f)|$.

Théorème 6 (Principe d'addition des erreurs).

$$\mathsf{E}^{\mathtt{Qc}}(f) \le \sum_{i=0}^{n-1} \mathsf{E}^{\mathtt{Q}}_{[a_i,a_{i+1}]}(f). \tag{3.4}$$

3.2 Exemples fondamentaux : les formules composées du point milieu, du trapèze et de Simpson

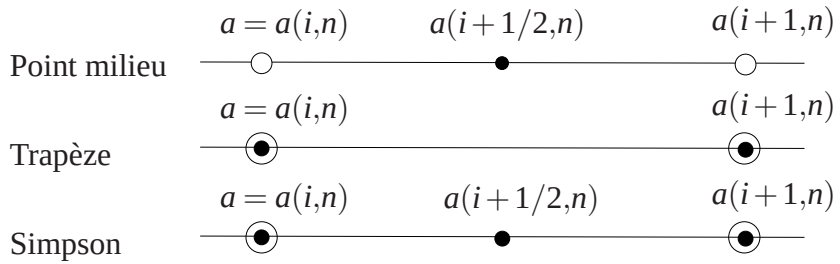
Soit $n \in \mathbb{N}^\star$, $I = [a,b]$ et $f \in \mathsf{C}[a,b]$. On pose $h(n) = (b-a)/n$ et $a(i,n) = a + ih(n)$.

3.3 Exemples

L'exécution est très rapide : pour Simpson avec $n = 700$, 0.125 seconde.

PRINCIPALES FORMULES DE QUADRATURES COMPOSÉES

$$I=[a,b],\ h(n)=(b-a)/n,\ a(i,n)=a+ih(n),\ f\in \mathsf{C}(I)$$



	Formule : $\mathtt{Qc}(f)$	Erreur : $\mathsf{E}^{\mathtt{Qc}}(f)$	Type de fonctions
Point milieu	$h(n)\cdot\sum_{i=0}^{n-1} f(a(i+1/2,n))$	$\frac{(b-a)^3}{24n^2}\cdot\max_{[a,b]}\lvert f^{(2)}\rvert$	$(f\in\mathsf{C}^2(I))$
Trapèze	$\frac{h(n)}{2}\cdot\left[f(a))+f(b)+2\sum_{i=1}^{n-1} f(a(i,n))\right]$	$\frac{(b-a)^3}{12n^2}\cdot\max_{[a,b]}\lvert f^{(2)}\rvert$	$(f\in\mathsf{C}^2(I))$
Simpson	$\frac{h(n)}{6}\left\{f(a)+f(b)+2\sum_{i=1}^{n-1} f(a(i,n))+4\sum_{i=0}^{n-1} f(a(i+1/2,n))\right\}$	$\frac{(b-a)^5}{2880n^4}\cdot\max_{[a,b]}\lvert f^{(4)}\rvert$	$(f\in\mathsf{C}^4(I))$

TAB. 1

n	Point milieu	Trapèze	Simpson
2.	3.1623529	3.1	3.1415686
4.	3.1468005	3.1311765	3.1415925
6.	3.1439074	3.1369631	3.1415926
8.	3.1428947	3.1389885	3.1415927
10.	3.142426	3.139926	3.1415927
920.	3.1415928	3.1415925	
930.	3.1415927	3.1415925	
1270.		3.1415926	

TAB. 2 – *comparasion des diverses méthodes pour* $\pi = \int_0^1 4/(1+x^2)dx$*=3,14159265358...*

n	Point milieu	Trapèze	Simpson
2.	- 0.0207603	0.0415927	0.0000240
4.	- 0.0052079	0.0104162	0.0000002
6.	- 0.0023148	0.0046296	1.328D-08
8.	- 0.0013021	0.0026042	2.365D-09
10.	- 0.0008333	0.0016667	6.200D-10
70.	- 0.0000170	0.0000340	5.329D-15
930.	- 9.635D-08	0.0000002	- 4.441D-16
2300.	- 1.575D-08	3.151D-08	4.441D-16

III

SOLUTIONS APPROCHÉES DES ÉQUATIONS

§ 1 INTRODUCTION

Soit $f : [a,b] \to \mathbb{R}$. Nous considérons l'équation $f(x) = 0$. Les problèmes que les numériciens doivent étudier sont les suivants.

(i) L'équation a-t-elle des solutions ?

(ii) Si oui, combien ?

(iii) Déterminer des valeurs aussi proches que l'on veut des solutions. (Sauf pour une classe restreinte de fonctions, on ne peut pas obtenir de solution exacte utilisable.)

Dans ce cours, les points (i) et (ii) ne seront pas abordés (sauf dans la dernière partie). Nous supposerons en général que l'équation $f(x) = 0$ admet une et une seule solution dans $[a,b]$. Parmi le grand nombre de méthodes disponibles, nous étudierons quatre techniques très classique pour résoudre le problème (iii).

(a) La méthode de **dichotomie**, basée sur le théorème des valeurs intermédiaires.

(b) Les méthodes de la **sécante** et de **Newton** qui consistent à remplacer l'équation $f(x) = 0$ par $p(x) = 0$ où p est un polynôme de degré 1 proche de f.

(c) La méthode dite du **point fixe** ou des **approximations successives**.

§ 2 MÉTHODE DE DICHOTOMIE (BISSECTION)

2.1 Définition

Soit f une fonction continue sur $[a,b]$. On suppose que

(i) f admet une et une seule racine dans $[a,b]$,

(ii) $f(a)f(b) < 0$.

Posons $c = \frac{a+b}{2}$. Si $f(c) = 0$, la racine est trouvée et le problème est résolu. Nous supposons que $f(c) \neq 0$.

- Si $f(a)f(c) < 0$ alors $f(a)$ et $f(c)$ sont de signes contraires donc la racine de f se trouve entre a et c car d'après le théorème des valeurs intermédiaires une fonction continue ne peut pas changer de signe sans s'annuler.
- Si, au contraire, $f(c)f(b) < 0$ alors la racine se trouve entre c et b.
- On recommence le processus en partant de $[a,c]$ au lieu de $[a,b]$ dans le premier cas, de $[c,b]$ au lieu de $[a,b]$ dans le second.

Algorithme 1. *Il construit trois suites $(a_n),(b_n)$ et (c_n) de la manière suivante*

(i) $a_1 = a$; $b_1 = b$.

(ii) Pour $n \geq 1$,

(a) $c_n = \frac{a_n+b_n}{2}$

(b) i. Si $f(c_n) = 0$ *alors* c_n *est la racine de* f *et le processus est arrêté*

ii. Sinon

- *Si* $f(c_n)f(b_n) < 0$ *alors* $a_{n+1} = c_n$ *et* $b_{n+1} = b_n$.
- *Si* $f(c_n)f(b_n) > 0$ *alors* $a_{n+1} = a_n$ *et* $b_{n+1} = c_n$.

L'algorithme ci-dessus s'appelle l'algorithme de **dichotomie** ou de **bissection**.

2.2 Etude de la convergence

Théorème 2. *Soit f continue sur $[a,b]$. Nous supposons que $f(a)f(b) < 0$ et que l'équation $f(x) = 0$ admet une et une seule solution r dans $[a,b]$. Si l'algorithme de dichotomie arrive jusqu'à l'étape $n+1$ (de sorte que $c_i \neq r$, $0 \leq i \leq n$) alors*

$$|r - c_{n+1}| \leq \frac{b-a}{2^{n+1}}.$$

Exemple 1. (i) L'équation $x^4 + x^3 - 1 = 0$ admet une solution (unique) dans $]0,1[$. Une approximation $\tilde{r}$ de la racine r avec une erreur moindre que 10^{-6} est obtenu en moins de 0.2 seconde : $\tilde{r} = 0.8191729$. La figure i donne les quatre premier termes de la suite.

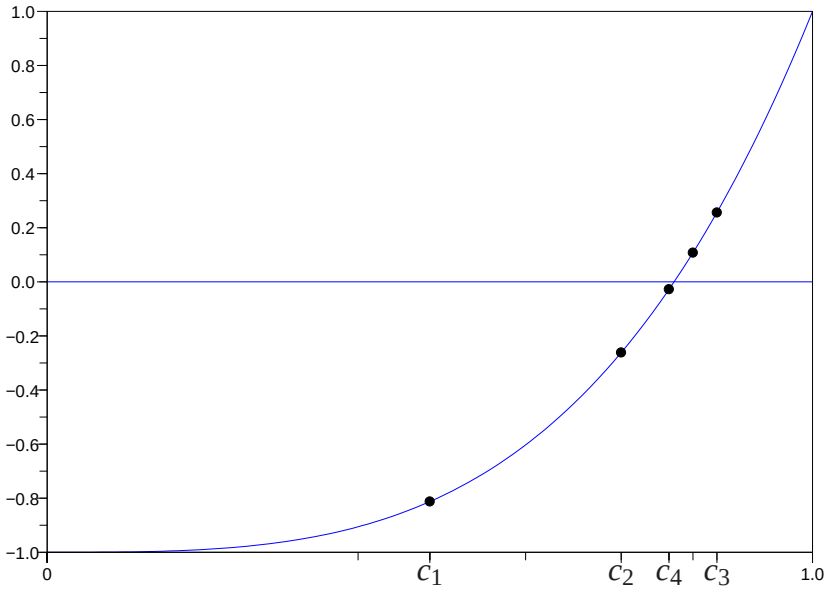


FIG. 1

(ii) L'équation $x - \sin x - 1/4 = 0$ admet une solution (unique) dans $]0,\pi/2[$. Une approximation $\tilde{r}$ de la racine r avec une erreur moindre que 10^{-6} est obtenu en moins de 0.2 seconde : $\tilde{r} = 1.1712288$. La table 1 donne un ensemble de valeurs.

n	c_n
1.	0.5
2.	0.75
3.	0.875
4.	0.8125
5.	0.84375
6.	0.828125
16.	0.8191681
17.	0.8191757
18.	0.8191719
19.	0.8191738
20.	0.8191729

n	c_n
1.	0.7853982
2.	1.1780972
3.	0.9817477
4.	1.0799225
5.	1.1290099
6.	1.1535536
16.	1.1712183
17.	1.1712303
18.	1.1712243
19.	1.1712273
20.	1.1712288

$x^4 + x^3 - 1 = 0,\ x \in [0,1]$ $\qquad$ $x - \sin x - 1/4 = 0,\ x \in [0,\pi/2]$

TAB. 1

§ 3 MÉTHODE DE NEWTON

3.1 Construction

Supposons que $f \in \mathrm{C}^1[a,b]$ et que l'équation $f(x) = 0$ admet une et une seule racine, notée r, dans $[a,b]$. L'idée de la **méthode de Newton** consiste à remplacer l'équation $f(x) = 0$ par l'équation $T_1(x) = 0$ où T_1 est un polynôme de Taylor de f de degré 1 en un point x_1 qu'il faudra bien choisir :

$$T_1(x) = f(x_1) + f'(x_1)(x - x_1).$$

L'équation $T_1(x) = 0$ admet pour racine

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

Il est naturel d'espérer que cette valeur sera proche de r i.e.

$$r \approx x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

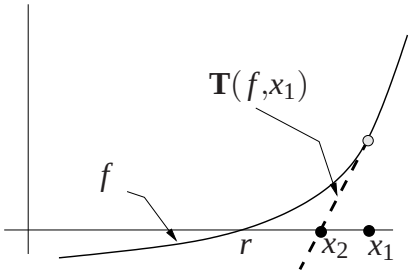


FIG. 2

Si $x_1 \in [a,b]$, nous pouvons itérer le procédé, en remplaçant $f(x) = 0$ par $T_1(x) = 0$ où

$$T_1(x) = f(x_1) + f'(x_2)(x - x_2)$$

dont la racine est

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)},$$

et

$$r \approx x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

Nous construisons ainsi par récurrence, *sous réserve que* $x_n \in [a,b]$, la suite

$$\begin{cases} x_1 &= b \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \geq 0) \end{cases} \qquad \textbf{(SCHÉMA DE NEWTON)}$$

E. 13. Donner (graphiquement) un exemple de fonction pour laquelle la suite x_n n'est pas définie (sort de l'ensemble de définition de la fonction).

3.2 Etude de la convergence

Nous devons répondre aux trois questions suivantes.

i) La suite (x_n) est-elle bien définie ?

ii) Si oui, converge-t-elle vers la racine r ?

iii) Si oui, quelle est la rapidité de convergence ?

Les réponses dépendent naturellement des propriétés de la fonction f considérée. De nombreux théorèmes apportent des réponses. Le suivant est l'un des plus simples. Ses hypothèses correspondent à la figure 2.

Théorème 3. *Soit f une fonction de classe C^2 sur un intervalle ouvert I contenant $[a,b]$ telle que f' et f'' soient strictement positives sur I (f est strictement*

croissante convexe). Nous supposons que $f(b) > 0$, $f(a) < 0$ et nous appelons r l'unique solution de l'équation $f(x) = 0$ dans $[a,b]$.

(i) La suite de Newton

$$\begin{cases} x_1 & = & b \\ x_{n+1} & = & x_n - \frac{f(x_n)}{f'(x_n)} \quad (n \geq 0) \end{cases}$$

est bien définie,

(ii) Elle converge vers r en décroissant,

(iii) L'estimation suivante est vraie

$$|x_n - r| \leq \frac{M_2}{2m_1}(x_n - r)^2$$

où $M_2 = \sup_{[a,b]} f''$ et $m_1 = \inf_{[a,b]} f'$.

Exemple 1. L'équation

$$3x^5 - x^4 - 1 = 0 \tag{3.1}$$

admet une et une seule racine r dans $[0,1]$. En effet la fonction $f(x) = 3x^5 - x^4 - 1$ a pour dérivée $f'(x) = x^3(15x-4)$ et, sur $[0,1]$ elle décroît de $f(0) = -1$ jusqu'à $f(4/15) \approx -1{,}001$ puis crôit jusqu'à $f(1) = 1$. En particulier $r \in]4/15,1[$. Par ailleurs, puisque $f''(x) = 12x^5(5x-1)$, f est strictement convexe sur $[1/5,1]$ en particulier sur $[4/15,1]$ puique $4/15 > 1/5$. Nous pouvons appliquer le théorème sur l'intervalle $[4/15,1]$ en prenant Comme point de départ $x_0 = 1$. Les dix premiers termes de la suite de Newton sont donnés dans le tableau 2. Remarquons que l'on obtient les six premières décimales de r dès le quatrième terme de la suite.

E. 14. Justifier l'emploi de la suite de Newton pour l'équation $x - \sin(x) - 1/4 = 0$.

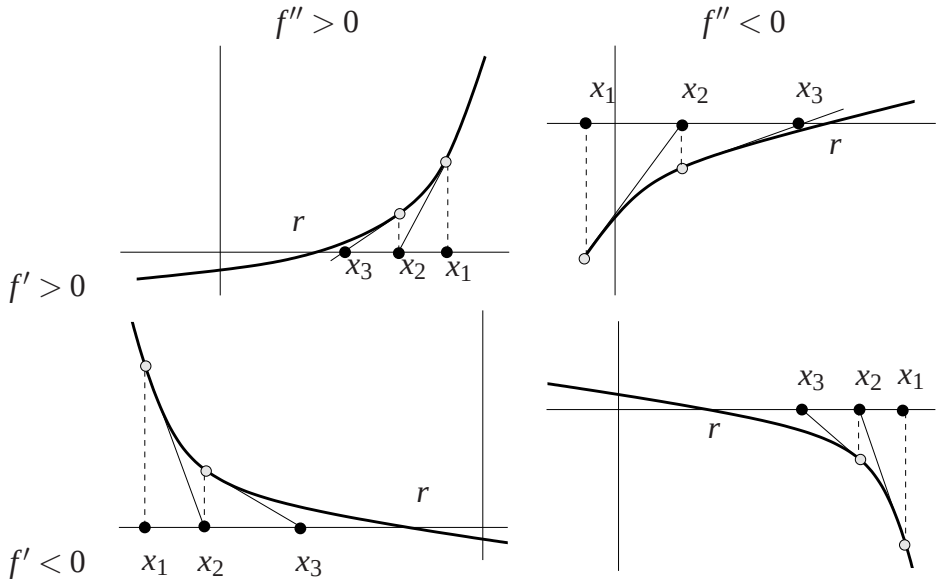
A cause de la relation $|x_{n+2} - r| \leq C(r - x_{n+1})^2$, on dit que la méthode de Newton est d'**ordre** 2. Une telle propriété implique une convergence très rapide. Par exemple, si, au rang n l'erreur est de l'ordre de 10^{-3}, au rang $n+1$, elle sera au pire de l'ordre de 10^{-6}, au rang $n+2$, 10^{-12}...

3.3 Autres versions

Il est facile d'adapter le théorème précédent pour traiter toutes les équations de la forme $f(x) = 0$ lorsque la fonction f et sa dérivée sont toutes deux strictement monotone. il y a quatre cas à considérer, ils sont donnés dans la figure 3.

$3x^5 - x^4 - 1 = 0$			$f(x) = x - \sin(x) - 1/4$		
n	x_n	$x_n - x_{n-1}$	n	x_n	$x_n - x_{n-1}$
1.	1.		1.	1.5707963 ($\pi/2$)	
2.	0.9090909	- 0.0909091	2.	1.25	- 0.3207963
3.	0.8842633	- 0.0248276	3.	1.1754899	- 0.0745101
4.	0.8826212	- 0.0016421	4.	1.1712433	- 0.0042467
5.	0.8826144	- 0.0000068	5.	1.1712297	- 0.0000136
6.	0.8826144	- 1.161D-10	6.	1.1712297	- 1.397D-10
7.	0.8826144	0.	7.	1.1712297	2.220D-16
8.	0.8826144	0.	8.	1.1712297	- 2.220D-16
9.	0.8826144	0.	9.	1.1712297	2.220D-16
10.	0.8826144	0.	10.	1.1712297	- 2.220D-16

TAB. 2 – *Premiers termes de deux suites de Newton*



TAB. 3 – *Quatre schémas de de Newton*

§ 4 MÉTHODE DE LA SÉCANTE

4.1 Construction

Supposons que $f \in \mathsf{C}[a,b]$, que l'équation $f(x) = 0$ admet une et une seule solution r dans $[a,b]$ et enfin que $f(a) < 0$, $f(b) > 0$. L'idée de la **méthode de sécante** consiste à remplacer l'équation $f(x) = 0$ par l'équation $\mathbf{L}[a,b;f](x) = 0$. Nous savons que

$$\mathbf{L}[a,b;f](x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$$

donc l'unique solution de $\mathbf{L}[a,b;f](x) = 0$ est donnée par

$$\begin{aligned} x_1 &= -f(a)\frac{b-a}{f(b)-f(a)} + a \\ &= \frac{-f(a)b + af(a) + af(b) - af(a)}{f(b)-f(a)} \\ &= \frac{af(b)-bf(a)}{f(b)-f(a)}. \end{aligned}$$

Comme pour la méthode de Newton, il est naturel d'espérer que cette valeur sera proche de r i.e.

$$r \approx x_1 = \frac{af(b)-bf(a)}{f(b)-f(a)}.$$

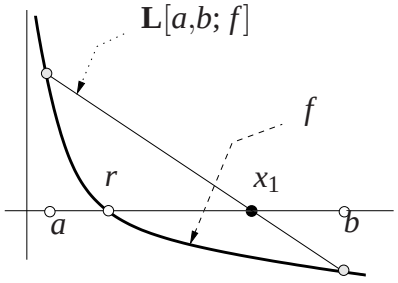


FIG. 3

Si $x_1 \in [a,b]$, le procédé peut être itéré en remplaçant $f(x) = 0$ par $\mathbf{L}[x_1,b;f](x) = 0$,autrement dit on fait jouer à x_1 le rôle que jouait précédemment a. nous pourrions évidemment envisager l'autre stratégie : garder a et remplacer b par x_1. Le choix de la stratégie, comme nous le verrons, est dicté par par la nature de la fonction f.

$$r \approx x_2 = \frac{x_1 f(b) - b f(x_1)}{f(b) - f(x_1)}.$$

En continuant, nous construisons par récurrence, *sous réserve* que $x_n \in [a,b]$, la suite

$$\begin{cases} x_0 &= a \\ x_{n+1} &= \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)} \quad (n \geq 0) \end{cases} \qquad \textbf{(SCHÉMA DE LA SÉCANTE)}$$

E. 15. Donner sur un exemple graphique une équation $f(x) = 0$ admettant une unique solution mais pour laquelle la suite de la sécante ne peut pas être construite.

4.2 Etude de la convergence

Nous devons répondre aux mêmes questions que pour la méthode de Newton. Les hypothèses du théorème suivant correspondent à la figure 4.1.

Théorème 4 (†)**.** *Soit f une fonction de classe C^2 sur un intervalle ouvert I contenant $[a,b]$ telle que f' et f'' soient strictement positives sur I (f est strictement croissante convexe). nous supposons que $f(b) > 0$, $f(a) < 0$ et nous notons r l'unique solution de l'équation $f(x) = 0$ dans l'intervalle $[a,b]$. A) La suite*

$$\begin{cases} x_0 &= a \\ x_{n+1} &= \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)} \quad (n \geq 0) \end{cases}$$

est bien définie. B) Elle converge en croissant vers r et C) nous avons l'estimation

$$|x_n - r| \leq \frac{M_2}{2m_1}(x_n - x_{n-1})(b - x_n)$$

où $M_2 = \sup_{[a,b]} f''$ et $m_1 = \inf_{[a,b]} f'$.

Exemple 1. Nous reprenons dans la table 4 les exemples étudiés ci-dessus avec la méthode de Newton.

E. 16. Indiquer comment adapter la méthode de la sécante à la résolution d'équations $f(x) = 0$ lorsque la fonction f et sa dérivée sont strictement monotones. On donnera un tableau correspondant au tableau 3 pour la méthode de Newton.

E. 17. Sous les hypothèses des deux théorèmes précédents (Th. 3 et Th. 4), on construit la suite $(\underline{x}_n)$ fournie par la méthode de la sécante et la suite $(\overline{x}_n)$ fournie par la méthode de Newton. Montrer que lorsque $\underline{x}_n$ et $\overline{x}_n$ ont les mêmes k premières décimales, ce sont aussi les k premières de r.

$3x^5 - x^4 - 1 = 0$ $\qquad$ $f(x) = x - \sin(x) - 1/4$

n	x_n	$x_n - x_{n-1}$
1.	0.15	- 1.
2.	0.5750592	0.4250592
3.	0.7787569	0.2036977
4.	0.8533380	0.0745812
5.	0.8749467	0.0216086
8.	0.8824870	0.0003733
9.	0.8825820	0.0000950
10.	0.8826062	0.0000242
11.	0.8826123	0.0000061
12.	0.8826139	0.0000016
13.	0.8826143	0.0000004
15.	0.8826144	2.570D-08
18.	0.8826144	4.226D-10

n	x_n	$x_n - x_{n-1}$
1.	0.15	
2.	1.4921931	1.3421931
3.	1.1336931	- 0.3585000
4.	1.1767653	0.0430721
5.	1.170437	- 0.0063282
6.	1.1713436	0.0009066
9.	1.1712293	- 0.0000027
10.	1.1712297	0.0000004
11.	1.1712296	- 5.563D-08
18.	1.1712297	7.039D-14
19.	1.1712297	- 9.992D-15
20.	1.1712297	1.332D-15

TAB. 4 – *Premiers termes de deux suites de méthode de la sécante*

§ 5 LA MÉTHODE DU POINT FIXE (DES APPROXIMATIONS SUCCESSIVES)

5.1 Introduction

Dans cette partie, nous considérons les équations de la forme $x = g(x)$. Nous étudierons un théorème qui, à la fois

(i) garantit l'existence et l'unicité de la solution

(ii) fournit une suite qui converge rapidement vers la solution.

Le procédé employé — les approximations successives — est un des plus fondamentaux de l'analyse. Il peut être généralisé à des équations plus compliquées dans lesquelles les inconnues sont des fonctions (par exemple les équations différentielles). Remarquons que les équations $f(x) = 0$ peuvent souvent être avantageusement mises sous la forme $x = g(x)$ (cf. exercices). Cependant en pratique la différence entre les deux types d'équations est plus grande qu'un simple différence de forme ne pourrait laisser penser.

5.2 Enoncé du théorème du point fixe

Théorème 5. *Soit I un intervalle fermé (non nécessairement borné) et g une fonction de I dans I. S'il existe un réel $k < 1$ tel que*

$$|g(x) - g(y)| \leq k|x - y| \quad \forall x,y \in I$$

alors l'équation

$$g(x) = x$$

admet une et une seule solution dans I. Cette solution est limite de la suite (x_n) définie par

$$\begin{cases} x_0 &= a \in I \\ x_{n+1} &= g(x_n) \quad (n \geq 0) \end{cases} \qquad \text{(SCH. DES APPROX. SUCC.)}.$$

(On est libre de choisir n'importe quel x_0 dans I). De plus, si s est la solution de l'équation $g(x) = x$ alors

$$|s - x_n| \leq \frac{k^n}{1-k}|x_1 - x_0| \qquad (n \geq 1).$$

L'intervalle I est de la forme $I = \mathbb{R}$ ou $I =]-\infty, a]$ ou $[a, +\infty[$ ou $[a,b]$. Il est essentiel que g prenne ses valeurs dans I c'est-à-dire que son ensemble image soit inclus dans son ensemble de définition, faute de quoi nous ne serions plus sûrs que la suite (x_n) soit bien définie.

Lorsqu'une fonction vérifie une inégalité

$$|g(x) - g(y)| \leq k|x - y| \quad \forall x,y \in I$$

avec $0 \leq k < 1$, on dit que f est **contractante** ou bien que c'est une **contraction** de constante k. Les fonctions contractantes sont continues en tout point. Fixons $x_0 \in I$ et montrons la continuité en x_0. Nous devons établir que $\forall \varepsilon > 0$, $\exists \eta > 0$ tel que les conditions ($|x - x_0| \leq \eta$ et $(x,y) \in I \times I$) impliquent $|g(x) - g(x_0)| \leq \varepsilon$. Il suffit de prendre $\eta = \varepsilon / k$.

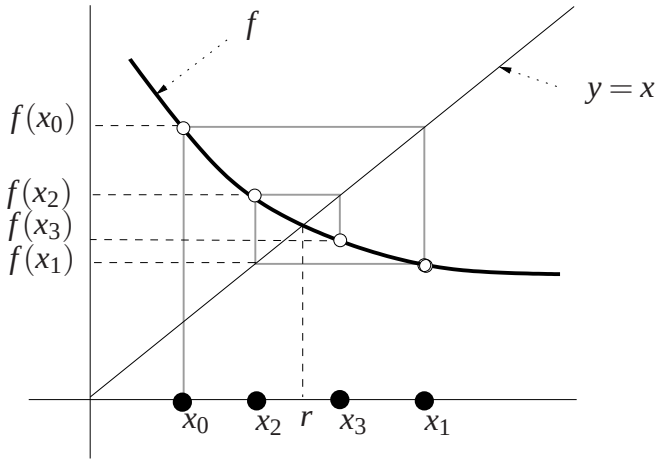
Lorsque g est dérivable, pour qu'elle soit contractante de constante k, il suffit que

$$\sup_I |g'| \leq k.$$

En effet d'après le théorème des accroissements finis,

$$|g(x) - g(y)| = |g'(c)||x - y| \leq k|x - y|.$$

Remarquons enfin que les suites définies par les schémas de Newton et de la sécante sont des cas particuliers de schémas d'approximations successives. Dans le premier cas on a $g(x) = x - \frac{f(x)}{f'(x)}$ et dans le second $g(x) = \frac{xf(b) - bf(x)}{f(b) - f(x)}$. Cependant, il n'est généralement pas aisé d'appliquer le théorème précédent dans ces cas particuliers.



Com. Construction des quatre premiers termes d'un suite d'approximation successive.

FIG. 4 – *Méthode du point fixe.*

5.3 Illustration graphique

La figure 4 montre un exemple de construction des permiers termes de la suite des approximations successives. On notera que les points $M_n = (x_n, f(x_n)$ convergent en "s'enroulant autour" de $(r, f(r)) = (r,r)$.

Exemple 1. La table 5 donne les résultats obtenus en appliquant la méthode du point fixe à l'équation $x = \sin(x) + 1/4$ en prenant deux points de départ différents.

E. 18. Montrer que la fonction f définie par $f(x) = \sin(x) + 1/4$ vérifie bien les conditions du théorème du points fixe en prenant comme intervalle de départ $I = [0,\pi/2]$.

5.4 Eléments de la démonstration du théorème du point fixe

$f(x) = x - \sin(x) - 1/4$			$f(x) = x - \sin(x) - 1/4$		
n	x_n	$x_n - x_{n-1}$	n	x_n	$x_n - x_{n-1}$
1.	1.		1.	0.5	
2.	1.091471	0.0914710	2.	0.7294255	0.2294255
3.	1.1373063	0.0458353	3.	0.9164415	0.1870159
4.	1.1575053	0.0201990	4.	1.0434407	0.1269993
5.	1.165804	0.0082987	5.	1.1141409	0.0707001
6.	1.1691054	0.0033014	6.	1.1475323	0.0333914
7.	1.1704012	0.0012958	7.	1.1617531	0.0142208
8.	1.1709071	0.0005058	8.	1.1675018	0.0057487
9.	1.1711041	0.0001971	9.	1.169773	0.0022713
10.	1.1711808	0.0000767	10.	1.170662	0.0008890
15.	1.1712292	0.0000007	15.	1.1712246	0.0000080
20.	1.1712296	6.090D-09	20.	1.1712296	7.084D-08
25.	1.1712297	5.426D-11	25.	1.1712297	6.312D-10
30.	1.1712297	4.834D-13	30.	1.1712297	5.624D-12

TAB. 5 – *Premiers termes d'une suite d'approximations successives*

IV

RÉSOLUTION DES SYSTÈMES LINÉAIRES. MÉTHODES DIRECTES.

§ 1 RAPPEL SUR LES SYSTÈMES LINÉAIRES

Tous les nombres considérés sont des réels mais tout ce qui sera dit dans ce chapitre reste vrai avec des nombres complexes*. nous considérons le système de n équations à n inconnues suivant

$$\begin{array}{c} \mathbf{L_1} \\ \mathbf{L_2} \\ \vdots \\ \mathbf{L_i} \\ \vdots \\ \mathbf{L_n} \end{array} \left\{ \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & c_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & c_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{i1}x_1 & + & a_{i2}x_2 & + & \dots & + & a_{in}x_n & = & c_i \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & c_n \end{array} \right. . \tag{1.1}$$

Les a_{ij} sont appelés les **coefficients** du système, les x_i, les inconnues (ou les solutions), les c_i forment le **second membre**. L'expression

$$\mathbf{L_i} \ : \ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = c_i \tag{1.2}$$

s'appelle la i-ième **ligne** du système. Le système (1.1) se représente aussi sous la

* Les nombres pourraient d'ailleurs être pris dans n'importe quel corps commutatif.

forme compacte

$$\sum_{j=1}^{n} a_{ij}x_j = c_i \qquad (i = 1,2,\ldots,n). \tag{1.3}$$

Au système linéaire ci-dessus est associée l'**équation matricielle**

$$AX = C \tag{1.4}$$

où la matrice A et les vecteurs X et C sont définis par

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \ldots & a_{in} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}, \qquad C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_i \\ \vdots \\ c_n \end{pmatrix} \tag{1.5}$$

Notons que dans a_{ij}, l'indice i désigne la ligne tandis que j désigne la colonne. Le vecteur colonne* X est appelé le **vecteur inconnu** (ou **vecteur solution**) et C est le **vecteur second membre**.

Rappelons en fait qu'une application linéaire $\mathscr{A}$ de $\mathbb{R}^n$ dans lui-même et canon uniquement associé à la matrice A. Cette application est définie par

$$\mathscr{A}(x) = \big(\sum_{j=1}^{n} a_{1,j}x_j, \ldots, \sum_{j=1}^{n} a_{i,j}x_j, \ldots, \sum_{j=1}^{n} a_{n,j}x_j\big), \quad x = (x_1,\ldots,x_n). \tag{1.6}$$

E. 19. Rappeler les liens entre les des images des éléments de la base canonique de $\mathbb{R}^n$ par l'application linéaire $\mathscr{A}$ et les coefficients de la matrice A.

Théorème 1 (**Règle fondamentale**, †)**.** *On ne modifie pas les solutions d'un système linéaire si on ajoute à une ligne une combinaison linéaire des autres lignes. On écrit*

$$\mathbf{L_i} \longleftarrow \mathbf{L_i} + \sum_{j\neq i} \alpha_j \mathbf{L_j}.$$

Il faut faire attention à ne pas oublier d'effectuer également la combinaison linéaire au niveau du second membre.

Théorème 2 (†)**.** *Pour que le système (1.1) admette une et une seule solution il faut et il suffit que* $\det A \neq 0$*. Dans ce cas la matrice A est inversible et l'unique solution est donnée par* $X = A^{-1}(C)$*.*

* Un vecteur $x \in \mathbb{R}^n$ est habituellement noté $x = (x_1,x_2,\ldots,x_n)$ mais pour des raisons propres au calcul matriciel, Lorsque nous lui appliquons une matrice A, nous avons intérêt à le représenter comme une colonne X. Dans la suite nous ne distinguerons plus, au niveau de la notation, le vecteur ligne x du vecteur colonne X.

Lorsque le système (1.1) admet une et une seule solution, nous disons que c'est un **système régulier**.

E. 20. Rappeler les règles de calcul du déterminant d'une matrice.

La condition sur le déterminant de A est à son tour équivalente à des propriétés naturelles de l'application linéaire $\mathscr{A}$. De manière précise, nous avons

$$\det A \neq 0 \iff \mathscr{A} \text{bijective} \iff \mathscr{A} \text{surjective} \iff \mathscr{A} \text{injective} \iff \ker \mathscr{A} = \{0\}. \tag{1.7}$$

Rappelons qu'ici l'équivalence entre bijective, surjective et injective est vraie uniquement parce que $\mathscr{A}$ est une application linéaire entre deux espaces vectoriels de même dimension.

E. 21. Donner un exemple d'application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^3$ qui soit injective (mais pas surjective). Donner un exemple d'application linéaire de $\mathbb{R}^3$ dans $\mathbb{R}^2$ surjective mais non injective.

Théorème 3 (**Formules de Kramer**, †)**.** *Lorsque* $\det A \neq 0$ *la coordonnée* x_j *de la solution* x *du système (1.1) est donnée par la formule*

$$x_j = \frac{1}{\det A} \begin{vmatrix} a_{11} & \dots & a_{1j-1} & c_1 & a_{1j+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j-1} & c_2 & a_{2j+1} & \dots & a_{2n} \\ \vdots & & & & & & \vdots \\ a_{n1} & \dots & a_{nj-1} & c_n & a_{nj+1} & \dots & a_{nn} \end{vmatrix} \quad (j = 1, \dots, n)$$

Pour obtenir x_j *on doit donc calculer le déterminant obtenu à partir de* A *en substituant le vecteur* C *à la* j*-ième colonne de* A*.*

Malheureusement les formules de Kramer nécessitent un nombre d'opérations trop grand et elles sont en pratique inutilisables (cf. exercices). Il faut donc rechercher d'autres méthodes.

§ 2 LE CAS DES SYSTÈMES TRIANGULAIRES

Exception faite des systèmes diagonaux (ceux dont la matrice est une matrice diagonale) qui se réduisent à n équations du type $a_{ii}x_i = c_i$, $i = 1, \dots, n$, les systèmes les plus simples sont les **systèmes triangulaires**. Dans cette partie, nous donnons les algorithmes élémentaires pour résoudre les systèmes linéaires triangulaires par substitutions successives et étudions la complexité de ces les algorithmes. Nous verrons ensuite comment n'importe quel système régulier peut se réduire — autrement dit, est équivalent — à un système triangulaire.

Considérons les deux systèmes linéaires suivants.

$$(S_1) \qquad\qquad (S_2)$$

$$\begin{cases} l_{11}x_1 & & & & & = & c_1 \\ l_{21}x_1 & + & l_{22}x_2 & & & = & c_2 \\ l_{31}x_1 & + & l_{32}x_2 & + & l_{33}x_3 & = & c_3 \end{cases} \qquad \begin{cases} u_{11}x_1 & + & u_{12}x_2 & + & u_{13}x_3 & = & c_1 \\ & & u_{22}x_2 & + & u_{23}x_3 & = & c_2 \\ & & & & u_{33}x_3 & = & c_3 \end{cases}$$

(Système triangulaire inférieur) (Système triangulaire supérieur)

Chacun des deux systèmes se résout facilement par **substitutions successives** (à condition que les éléments diagonaux soient non nuls)

$$(S_1) \qquad\qquad (S_2)$$

$$\begin{cases} x_1 = \frac{c_1}{l_{11}} \\ x_2 = (c_2 - l_{21}x_1)/l_{22} \\ x_3 = (c_3 - l_{31}x_1 - l_{32}x_2)/l_{33} \end{cases} \qquad \begin{cases} x_3 = \frac{c_3}{u_{33}} \\ x_2 = (c_2 - u_{23}x_3)/u_{22} \\ x_1 = (c_1 - u_{12}x_2 - u_{13}x_3)/u_{33} \end{cases}$$

La technique de substitutions successives s'applique de la même manière aux systèmes triangulaires de n équations.

Algorithme 4 (Substitutions successives L)**.** *Les solutions du système triangulaire* inférieur

$$LX = C \qquad avec \quad \begin{cases} l_{ij} = 0 \quad si \quad i < j \\ l_{ii} \neq 0 \end{cases}$$

sont données par les relations

$$\begin{cases} x_1 = \frac{c_1}{l_{11}} \\ x_i = \frac{1}{l_{ii}} \left(c_i - \sum_{j=1}^{i-1} l_{ij}x_j \right) \end{cases} \qquad (i = 2,3,\ldots,n).$$

Algorithme 5 (Substitutions successives U)**.** *Les solutions du système triangulaire* supérieur

$$UX = C \qquad avec \quad \begin{cases} u_{ij} = 0 \quad si \quad i > j \\ u_{ii} \neq 0 \end{cases}$$

sont données par les relations

$$\begin{cases} x_n = \frac{c_n}{u_{nn}} \\ x_i = \frac{1}{u_{ii}} \left(c_i - \sum_{j=i+1}^{n} u_{ij}x_j \right) \end{cases} \qquad (i = n-1, n-2, \ldots, 1).$$

Théorème 6. *La résolution d'un système linéaire triangulaire (supérieur ou inférieur) de n équations à n inconnues par la méthode des substitutions successives nécessite n^2 opérations élémentaires.*

§ 3 L'ALGORITHME DE GAUSS

3.1 Description de l'algorithme dans le cas d'un système de 3 équations à 3 inconnues. Notion de pivot

Soit à résoudre le système suivant dont on suppose qu'il admet une et une seule solution (le déterminant de la matrice associée est donc supposé non nul)

$$S^{(0)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(0)}\\ \mathbf{L}_3^{(0)}\end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = c_3 \end{cases}.$$

Etape 1. Elimination de x_1 dans $\mathbf{L}_2^{(0)}$ et $\mathbf{L}_3^{(0)}$.

$$\left\| \begin{matrix} \mathbf{L}_2^{(0)} \longleftarrow \mathbf{L}_2^{(0)} - \frac{a_{21}}{a_{11}}\mathbf{L}_1^{(0)} \\ \mathbf{L}_3^{(0)} \longleftarrow \mathbf{L}_3^{(0)} - \frac{a_{31}}{a_{11}}\mathbf{L}_1^{(0)} \end{matrix} \right. .$$

Attention, on divise par a_{11}. Cela n'est donc possible que si a_{11} est non nul. On arrive à

$$S^{(1)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(1)}\end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + (a_{22} - \frac{a_{21}}{a_{11}}a_{12})x_2 + (a_{23} - \frac{a_{21}}{a_{11}}a_{13})x_3 = c_2 - \frac{a_{21}}{a_{11}}c_1 \\ 0 + (a_{32} - \frac{a_{31}}{a_{11}}a_{12})x_2 + (a_{33} - \frac{a_{31}}{a_{11}}a_{13})x_3 = c_3 - \frac{a_{31}}{a_{11}}c_1 \end{cases} .$$

que l'on écrit encore sous la forme

$$S^{(1)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(1)}\end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 = c_3^{(1)} \end{cases} .$$

Etape 2. Elimination de x_2 dans $\mathbf{L}_3^{(2)}$.

$$\left\| \mathbf{L}_3^{(1)} \longleftarrow \mathbf{L}_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}\mathbf{L}_2^{(1)} \right. .$$

Attention, on divise par $a_{22}^{(1)}$. Cela n'est donc possible que si $a_{22}^{(1)}$ est non nul. On arrive à

$$S^{(2)} \quad \begin{matrix}\mathbf{L}_1^{(0)}\\ \mathbf{L}_2^{(1)}\\ \mathbf{L}_3^{(2)}\end{matrix} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = c_1 \\ 0 + a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 = c_2^{(1)} \\ 0 + 0 + (a_{33}^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}a_{23}^{(1)})x_3 = c_3^{(1)} - \frac{a_{32}^{(1)}}{a_{22}^{(1)}}c_2^{(1)} \end{cases}$$

que l'on écrit encore sous la forme

$$S^{(2)} \quad \begin{matrix} \mathbf{L}_1^{(0)} \\ \mathbf{L}_2^{(1)} \\ \mathbf{L}_3^{(1)} \end{matrix} \quad \begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ 0 & + & a_{22}^{(1)}x_2 & + & a_{23}^{(1)}x_3 & = & c_2^{(1)} \\ 0 & + & 0 & + & a_{33}^{(2)}x_3 & = & c_3^{(2)} \end{cases}.$$

Ce dernier système est triangulaire supérieur, on peut donc le résoudre rapidement par substitutions successives comme expliqué dans la partie précédente. Il reste à examiner si, et comment, on peut modifier la méthode dans le cas où un des nombres par lesquels on doit diviser s'avère être égal à 0. Supposons que $a_{11} = 0$. Nous avons alors $a_{21} \neq 0$ ou $a_{31} \neq 0$ sinon la première colonne de la matrice du système serait nulle et son déterminant vaudrait 0 ce qui est contraire à l'hypothèse. Supposons pour fixer les idées que $a_{22} \neq 0$, nous permutons alors les lignes $\mathbf{L}_1^{(0)}$ et $\mathbf{L}_2^{(0)}$ et commençons la méthode décrite ci-dessus à partir du système

$$\begin{cases} a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & c_2 \\ 0 & + & a_{12}x_2 & + & a_{13}x_3 & = & c_1 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & = & c_3 \end{cases}.$$

Dans la deuxième étape, si nécessaire, c'est-à-dire si $a_{22}^{(1)} = 0$, nous pouvons permuter les lignes $\mathbf{L}_2^{(1)}$ et $\mathbf{L}_3^{(1)}$ de telle sorte que nous diviserons à nouveau par un nombre non nul.

Les nombres par lesquels on divise dans les diverses étapes de l'algorithme s'appellent les **pivots de Gauss**. Pour que l'algorithme fonctionne, ces nombres doivent être non nuls. Cependant, pour la précision des calculs, on a intérêt à les choisir les plus grands possibles en valeur absolue. Cette question ne sera pas abordée ici (cf. exercices).

Dans la partie suivante on décrit l'algorithme de Gauss ci-dessus dans le cas d'un système à n équations et n inconnues. On l'énonce dans le cas où la matrice du système n'est pas supposée inversible en le munissant d'une instruction d'arrêt pour le cas où $\det A = 0$.

3.2 Algorithme de Gauss (sans optimisation de pivot)

Algorithme 7 (Notation Ligne)**.** *On considère le système linéaire de matrice associée A*

$$\left(\mathbf{L_k}: \qquad \sum_{j=1}^{n} a_{kj}x_j = c_k, \qquad k = 1,2,\ldots,n\right)$$

1 *Pour* $j = 1,\ldots,n-1$ *faire sauf ordre d'arrêt*

1.1 **Si** $a_{ij} = 0$ *pour tout* $i \geq j$ *alors ARRÊT. (A est non inversible.)*

Sinon *soit* $i_o = \inf\{i, a_{ij} \neq 0\}$, *faire*

$$\mathbf{L_j} \longleftarrow \mathbf{L_{i_0}}$$
$$\mathbf{L_{i_0}} \longleftarrow \mathbf{L_j}$$

1.2 *Pour* $i > j$ *faire*

$$\mathbf{L_i} \longleftarrow \mathbf{L_i} - \frac{a_{ij}}{a_{jj}}\mathbf{L_j}.$$

2 *Résoudre le système (triangulaire) formé des (nouvelles) lignes* L_i *par la méthode des substitutions successives.*

Algorithme 8 (Notation Coefficient)**.** *On considère le système linéaire de matrice associée A*

$$\Big(\mathbf{L_k}: \qquad \sum_{j=1}^{n} a_{kj}x_j = c_k, \qquad k = 1,2,\ldots,n\Big)$$

1 *Pour* $j = 1,\ldots,n-1$ *faire sauf ordre d'arrêt*

1.1 **Si** $a_{ij} = 0$ *pour tout* $i \geq j$ *alors ARRÊT. (A est non inversible.)*
Sinon *soit* $i_o = \inf\{i, a_{ij} \neq 0\}$, *faire*

$$\begin{aligned} a_{jl} &\longleftarrow a_{i_0 l} \quad \textit{pour} \quad l \geq j \\ a_{i_0 l} &\longleftarrow a_{jl} \quad \textit{pour} \quad l \geq j \\ c_j &\longleftarrow c_{i_0} \\ c_{i_0} &\longleftarrow c_j \end{aligned}$$

1.2 *Pour* $i > j$ *faire*

1.2.1

$$m_i = \frac{a_{ij}}{a_{jj}}$$
$$c_i \longleftarrow c_i - m_i c_j.$$

1.2.2 *pour* $k > j$ *faire*

$$a_{ik} \longleftarrow a_{ik} - m_i a_{jk}.$$

2 *Résoudre le système (triangulaire)*

$$\Big(\sum_{i=k}^{n} a_{ki}x_i = b_k \qquad k = 1,2,\ldots,n\Big)$$

3.3 Coût de l'algorithme de Gauss

Théorème 9. *Le nombre N_n d'opérations élémentaires nécessaires pour résoudre un système linéaire à n équations et n inconnues (de déterminant non nul) par la méthode de Gauss est asymptotiquement égal à $2n^3/3$. On écrit $N_n \sim 2n^3/3$ et cela signifie* $\lim_{n\to\infty} \frac{N_n}{2n^3/3} = 1$.

INDEX